Project Manager's Guide to Managing Impact and Process Evaluation Studies

Prepared for:
Office of Energy Efficiency and Renewable Energy (EERE)
Department of Energy

Prepared by:
Yaw O. Agyeman, Lawrence Berkeley Laboratory
&
Harley Barnes, Lockheed Martin

August 2015

U.S. Department of Energy
Energy Efficiency and Renewable Energy
Bringing you a prosperous future where energy is clean, abundant, reliable, and affordable

Acknowledgments

This "Project Manager's Guide to Managing Impact and Process Evaluation Studies," was completed for the U.S. Department of Energy (DOE) by Lawrence Berkeley National Laboratories (LBNL), Berkeley, California, U.S.A. under contract number EDDT06 and subcontract number 7078427.

Yaw Agyeman, LBNL, and Harley Barnes, Lockheed Martin, were the authors for the guide. Jeff Dowd, DOE's Office of Energy Efficiency and Renewable Energy (EERE), Office of Strategic Programs, was the DOE Project Manager.

EERE internal reviewers were:

- Adam Cohen, EERE
- Craig Connelly, EERE
- Michael Li, EERE
- John Mayernik, NREL

External peer reviewers included:

- Gretchen Jordan, 360 Innovation, LLC
- Ken Keating, Consultant

An earlier 2006 guide, "EERE Guide for Managing General Program Evaluation Studies", provided the conceptual foundations for this guidance document. Harley Barnes co-authored the earlier guide with Gretchen Jordan, Founder & Principal, 360 Innovation LLC (formerly technical staff with Sandia National Laboratories).

Notice

Table of Contents

1.0 INTRODUCTION

1.1 Purpose and Scope

Myriad directives from the White House have emphasized accountability and evidence-based decision-making as key priorities for the federal government, bringing renewed focus to the need for evaluative activities across federal agencies.[1] The U.S. Department of Energy's (DOE) Office of Energy Efficiency and Renewable Energy has responded to these directives positively, through a systemic approach of capacity-building (to which this guide contributes), standard setting, and commissioning of evaluation studies.

The purpose of this Guide is to help managers of EERE evaluation projects create and manage objective, high quality, independent, and useful impact and process evaluations.[2] The step-by-step approach described in this Guide is targeted primarily towards program staff with responsibility for planning and managing evaluation projects for their office, but who may not have prior training or experience in program evaluation. The objective is to facilitate the planning, management, and use of evaluations, by providing information to help with the following:

- Determine why, what and when to evaluate
- Identify the questions that need to be answered in an evaluation study
- Specify the type of evaluation(s) needed
- Hire a qualified independent third-party evaluator
- Monitor the progress of the evaluation study
- Implement credible quality assurance (QA) protocols
- Ensure the evaluation report presents accurate and useful findings and recommendations
- Ensure that the findings get to those who need them
- Ensure findings are put to appropriate use.

1.2 What is Program Evaluation?

Program evaluations are systematic and objective studies, conducted periodically or on an ad hoc basis, to assess how well a program is achieving its intended goals. A program evaluation study is a management tool that answers a broader range of critical questions about program improvement and accountability than regular performance monitoring and reporting activities.[3] Program *performance monitoring* and reporting provide information on performance and output achievement. Program *evaluation* provides answers to questions about effects in the population of interest that occurred because of the program rather than because of other influences (impact evaluation), and to questions about the efficiency and effectiveness of the program implementation processes (process evaluation).

---

[1] The list of pertinent memoranda includes: OMB Memo M-13-17 (encourages federal agencies to use evidence and innovation to improve budget submissions and performance plans); OMB Circular A-11 Section 51.9 (emphasizes that OMB will evaluation budget submissions based in part on use of evidence in shaping resource allocations); OMB M-12-14 (focuses on use of evidence and evaluation in 2014 budget); an OMB M-10-01 (points to increased emphasis on program evaluations).

[2] An evaluation project manager is a staff member with responsibility for planning, commissioning, managing and facilitating the use of impact and process evaluation studies of EERE programs.

[3] Office of Management and Budget, "Preparation and Submission of Strategic Plans, Annual Performance Plans, and Annual Program Performance Reports." OMB Circular, No. A-11 (2002), Part 6, Section 200.2.

The focus of this Guide is on impact and process (also known as implementation) evaluations performed by outside experts and independent third-party evaluators.[4] The relevant types are described in the box below. These types of evaluations have either a retrospective or contemporary focus, with a view to assessing past or current performance and achievements, and developing recommendations for improvements. Evaluations investigate what works and why; impact evaluations provide evidence that outcomes have occurred, and some portion of those outcomes can be attributed to the program. Program evaluations require levels of detail in data collection and analyses that go beyond routine performance monitoring and reporting. Program evaluations can help technology or deployment managers and office directors (henceforth referred to as "managers") determine where and when to invest, what kinds of timely adjustments may be needed, and whether an investment was worth the effort.

Types of Program Evaluations that are the Focus of this Guide

Process or Implementation Evaluations – Evaluations that examine the efficiency and effectiveness of program implementation processes. The results of the evaluation help managers decide how to improve program operations, design, or targeting.[5]

Impact Evaluations – Evaluations that provide evidence that outcomes have occurred, and estimate the proportion(s) of the outcome(s) that are attributable to the program rather than to other influences. These findings demonstrate the value of the program investment to key stakeholders and, if designed to do so, help managers decide whether to continue the program, and at what level of effort.

Cost-benefit / Cost-effectiveness Evaluations – A form of impact evaluation that analyzes and calculates quantitative economic benefits, and compares benefits attributable to the program to the program's costs. Cost-benefit evaluations show, in monetary units, the relationship between the value of the outcomes of a program and the costs incurred to achieve those benefits. Cost-effectiveness evaluations are similar, but the benefits are not rendered in monetary units. Combined with the other evaluations, cost-benefit and cost-effectiveness findings help managers justify past investments and decide on future investments.[6]

A later section of this Guide discusses the strength of an evaluation's results. A manager anticipating a need to rate the strength of an evaluation's results may want to assess the ability of one of these evaluations to provide strong evidence of a program's effectiveness before the evaluation is initiated. Such a pre-study assessment is called an *evaluability assessment*. An evaluability assessment is usually a relatively low-cost early subjective look at whether the methods and resources available can produce evaluation results having the strength needed to make them useful to a program's stakeholders. This Guide will discuss evaluability assessments in Section 4.

1.3 Why, What and When to Perform Evaluations

Evaluations serve programs in two critical ways – program improvement and accountability. Impact evaluations are motivated primarily by the need for accountability – to demonstrate value

[4] Peer review of program or subprogram portfolios by independent external experts is a form of process evaluation.
[5] A process evaluation is sometimes called a "formative evaluation," and an impact evaluation is sometimes called a "summative evaluation." These terms, used primarily in the academic literature, are mostly omitted from this guide.
[6] Another type of evaluation, "Needs Assessment or Market Assessment," that involves assessing such things as customer needs, target markets, market baselines, barriers to adoption of energy efficiency and renewable energy, and how best to address these issues by the program in question, is not addressed explicitly in this Guide, although the principles are similar.

to key stakeholders – but also the desire for continuous improvement. Many evaluations are designed to serve both of these purposes.

- **Improvement:** Program impact (if designed to do so) and process evaluations help managers determine how well their programs are working by assessing the extent to which desired outcomes are being achieved and by identifying whether process improvements are needed to increase efficiency and effectiveness with respect to objectives. Program evaluation studies help managers proactively optimize their programs' performance.

- **Accountability:** Program impact and process evaluations also help managers and others demonstrate accountability for the use of public resources. Accountability includes the communication of fiscal responsibility and program value through reporting and targeted communication to key stakeholders.

In terms of what to evaluate, not every program, or part of a program, needs an impact evaluation. Some programs may be judged on monitored operational performance metrics only. Decisions on what to evaluate must consider the following factors:

- The investment is a priority for key stakeholders (e.g., White House / Congress / DOE Secretary or EERE Assistant Secretary);

- The size of the portfolio is substantial (e.g., the investment represents a significant proportion of total annual office budget);

- The program, subprogram or portfolio is a high profile one that has never been evaluated;

- The investment is of critical path importance to achieving office or EERE goals;

- Market penetration, a key intermediate outcome, might be occurring, but evidence is lacking;

- A prior evaluation for the program, subprogram or portfolio, need to be updated;

- There is interest in scaling up, down, or replicating the investment; or

- It is necessary to determine why an investment is not achieving intended results.

Developing a long-term evaluation strategy, with a schedule of planned and appropriately sequenced evaluation studies to meet learning and accountability needs, would enable the program to maximize its efficiency and effectiveness in the conduct of evaluations to maximize program success.

With regards to the timing of evaluations, there are no hard and fast rules on precisely when to conduct a program evaluation, except for ensuring that the evaluation results would be obtained in time for the decisions for which they are needed. However, over the program lifecycle, there are specific types of evaluations suitable for certain program phases and for which some general guidelines on frequency are advised. Table 1-1 presents periods of a program's life cycle and which impact and process evaluation is most appropriate to use.

Table 1-1. Guidance on Types and Timing of Program Evaluations

Program Life Cycle Stage	Type of Evaluation
Planning or early implementation	**Needs assessment:** Appropriate during program initiation and early implementation phase. These assessments can inform program strategies such as targeting, potential partnerships, and timing of investments. It is also the time to plan and instate, based on the program theory of change[7], data collection protocols to collect routine data for performance monitoring and impact evaluation. ***NOTE: Needs assessments are a special type of evaluation. This guide does not focus on this type of evaluation***.
During program operations	**Process evaluation:** Advisable once every 2-3 years, or whenever a need exists to assess the efficiency and effectiveness of the program's operations and barriers to its progress. Process evaluations can also be performed at any time to answer ad hoc questions regarding program operations. If results from consecutive evaluations of certain processes do not change, and the program context has not changed, subsequent evaluation of those processes can be performed less frequently. **Impact evaluation:** Suggested once every 3-5 years or annually if desired outcomes occur in that time frame. Results have multiple uses, including support of annual Government Performance and Results Act (GPRA) benefits analysis, budgeting, accountability and design improvements. An impact evaluation may be preceded by an evaluability assessment. **Cost-benefit evaluation:** Suggested once every 3-5 years. A cost-benefit evaluation is a special type of impact evaluation, with a focus on comparing benefits and costs of an intervention. It can be done separately, or as part of a broader impact evaluation.
Closeout or after end of program	**Process and impact evaluations after the program has ended:** Suggested timeframe is within one year of the end of the program, or after 5 years or more to follow up on some desired outcomes. Apply process evaluation lessons to the design of next-generation programs; use impact evaluation, including a cost-benefit evaluation if desired.

Depending on the intended uses of an evaluation, a manager may plan on a sequence of evaluations for each stage of a program life cycle, to be carried out over a time span consistent with the need for results to support particular decisions.

For example, process evaluations might be planned for, at scheduled intervals, to ensure that program implementation is proceeding according to plan, and successfully generating expected outputs, in conformance with stakeholder expectations and program objectives. Impact evaluations can also be planned for, to be undertaken when program activities are ready to be evaluated, with an eye on quantifying achieved impact and on how the results could be used for program improvement and for accountability.

[7] Theories of change aim to link activities to outcomes, to explain how and why a desired change can be reasonably expected from a particular intervention. It may be the case that empirical evidence has not yet been established regarding the sequence of expected transitions leading from intervention activity to desired outcomes. The theory of change then functions as a form of guide for hypothesis testing. Logic models might conceptually be viewed the *de facto* understanding of how program components are functioning, as a graphic illustration of the underlying program theory of change.

1.4 Overview of Steps, Roles, and Responsibilities

The Office of Management and Budget (OMB) and Congress require transparency and objectivity in the conduct of impact evaluations. To satisfy these requirements managers need to solicit independent evaluation experts to perform the evaluation studies described in this Guide.

Program managers will need to clearly define and formulate the evaluation objectives and expectations before selecting a qualified independent third-party evaluator. For this reason, it is important that the evaluation program managers, or the program staff assigned responsibility for an evaluation project, know all of the steps in this Guide. Familiarity with the steps involved in the conduct of a typical program evaluation and with evaluation terminology will facilitate communication with the independent evaluation experts who perform the evaluation.

> The steps in this Guide appear in the order in which they are often performed in practice. However, as with all processes of research and inquiry, most of the steps are iterative in execution and involve feedback loops.
>
> The steps are not prescriptive, but they *do* represent common practice for evaluations. In that sense, it will be valuable to review this Guide in its entirety and become familiar with its concepts before beginning to plan and formulate an evaluation.

This Guide divides the planning and management process for a program evaluation into seven major steps and describes briefly what each step entails. Table 1-2 presents these steps, matched to the roles and responsibilities of involved parties. Although the steps are listed as discrete events, in practice some of them overlap and are performed concurrently or interact with each other through feedback loops. That is to say, the evaluation management process is an iterative process, but the steps identified are essential elements of the process.

Although some of the steps listed in Table 1-2 need to occur in sequence, there is considerable iteration, especially for activities within the same step. For example, the activities in Step 1 will probably be performed not just iteratively but concurrently, to ensure that the different elements are in continuous alignment. The manager may then need to revisit Step 1 and seek expert advice while developing the statement of work (SOW) (Step 2) because change in one part affects other parts, as might occur when resource considerations invariably affect the choice of evaluation method.

After the independent third-party evaluator is hired, he or she will revisit Steps 1 and 2 to develop the details of the work described. However, regardless of the actual order in which the steps are performed, the uses and objectives of the study must be established (Step 1) before specifying the questions the evaluation must answer (Step 3). The next section offers some basic guidelines for the steps enumerated in Table 1-2.

Table 1-2. Steps, Roles, and Responsibilities for Performing and Managing Evaluation Studies

Steps in Performing and Managing Evaluation Studies	Roles and Responsibilities	
	DOE Evaluation Project Manager	Third Party Evaluator
Step 1. Prepare for the Evaluation		
• Initial Evaluation Planning (may be done in consultation with experts)		
○ Determine and prioritize the intended uses of evaluation information	✓	
○ Identify what kinds of evaluation information is needed for the intended uses and decide on the type of evaluation needed to develop the information	✓	
○ Align timeline for completing the evaluation with when information is needed	✓	
○ Determine the level of evaluation rigor needed to satisfy the intended uses of the results	✓	
○ Formulate an initial program logic model, metrics, and evaluation questions	✓	
○ Estimate evaluation cost and other resources needed	✓	
○ Organize background data and program records for use in the evaluation	✓	
Step 2. Hire an Independent Outside Evaluator		
• Develop the request for proposals (RFP)	✓	
• Implement the RFP competitive solicitation process to hire an independent evaluator	✓	
• Ensure EERE quality assurance protocol for the evaluation is set up to be implemented (i.e., a procedure for external peer review)	✓	
Step 3. Develop the Evaluation Plan		
• Develop a final program logic model, metrics, and researchable evaluation questions		✓
• Perform an evaluability assessment		✓
• Determine an appropriate research design		✓
• Establish a data collection plan		✓
• Choose the appropriate analytical method(s) for the selected research design	✓	✓
• Participate in peer review of the evaluation plan	✓	✓
Step 4. Conduct the Evaluation		
• Perform sampling, data collection, measurement and verification		✓
• Complete data analyses and calculations		✓
• Identify key findings		✓
Step 5. Manage the Evaluation Project During Implementation		
• Hold and participate in periodic project progress-review meetings	✓	✓
• Review project status reports from the third party evaluator	✓	✓
• Monitor evaluator's achievement of milestones and expenditures	✓	✓
• Manage the internal and external review process	✓	
• Anticipate and address technical and management challenges	✓	✓
Step 6. Report the Evaluation Results		
• Prepare draft and final evaluation reports using DOE reporting guidelines		✓
• Participate in peer review of draft evaluation report and publish final report	✓	✓
Step 7. Use the Evaluation Findings		
• Distribute the evaluation report and results	✓	
• Use the results to make decisions about the program	✓	
• Use the results for high impact communications	✓	
• Establish/Update Program Records for use in future evaluations	✓	

1.5 Guide Roadmap

This Guide is divided into eight sections, including this introductory section. Sections 2 through 8 provide guidance for the key steps involved in planning and managing an impact or process evaluation. Under each step, there are specific sub-steps that represent the tangible actions for the evaluation project manager and evaluation independent third-party evaluator.

Section 1. Introduction
Section 2. Step 1: Prepare for the Evaluation
Section 3. Step 2: Hire an Independent Outside Evaluator
Section 4. Step 3: Develop an Evaluation Plan
Section 5. Step 4: Conduct the Evaluation
Section 6. Step 5: Manage the Evaluation Project During Implementation
Section 7. Step 6: Report the Evaluation Findings
Section 8. Step 7: Use the Evaluation Results

The appendices contain examples of documents required at several steps in the evaluation process and related information.

Appendix A. Example Statement of Work for an R&D Evaluation Study
Appendix B. Example SOW for Non-R&D Evaluation Study
Appendix C. Example of a Request for Proposal (RFP) for a Program Evaluation Study
Appendix D. Procedures for Obtaining OMB Approval to Collect Information
Appendix E. Example of Non-R&D Evaluation Report Outline
Appendix F. Example of an R&D Evaluation Report Outline
Appendix G. Example of an Evaluation Study Peer Review Charter
Appendix H. Lessons Learned for Improving the Quality of EERE Evaluation Studies
Appendix I. Example of a Technical Evaluation Plan Outline
Appendix J. American Evaluation Association Ethical Principles for Evaluators
Appendix K. Program Evaluation Glossary

2.0 STEP 1. PREPARE FOR THE EVALUATION

This part of the Guide focuses on the essential steps to take in preparing for a program evaluation. The responsibility for these steps belongs to the program office. The DOE evaluation project manager and program office director must first determine why they need evaluation information. Once the need for, and intended uses of, evaluation information have been established, decisions can be made on which elements of the program must be evaluated, at what scope, within what timeframe, and the availability of needed data. From this, they can estimate the resource requirements for conducting the evaluation(s), and begin organizing internally to facilitate the conduct of the evaluation. Although this responsibility must be performed internally, the program office may choose to seek the early assistance of central office experts, or even an independent third-party evaluator, if needed. There are layers of technical knowledge necessary even in the preparation step.

2.1 Determine and Prioritize Intended Uses of Evaluation Information

The first step in preparing for an evaluation is to determine the uses of the evaluation data and prioritize among them if there are multiple needs. This, in turn, helps determine the evaluation objectives. In other words, evaluation objectives are determined by careful consideration of the possible decisions to which the evaluation's results will contribute. Some specific examples of decision types that a manager might take include:

- Continuing the program as is
- Expanding the program, consolidating components, or replicating components found to be most cost-effective
- Reallocating funding within the program; adding or reducing funding to the program
- Streamlining, refining, redesigning the program (e.g., to meet a pressing resource constraint)
- Setting more realistic objectives
- Discontinuing ineffective delivery components
- Discontinuing the program.

Each decision is strengthened by information from multiple sources such as impact and process evaluations, prospective data (forecasting), technology trends, market and policy data and analysis, and a manager's judgment and vision. The value-added of evaluation information for the decisions to be made must be taken into account. A clearly articulated set of intended uses, and a sense of the kinds of information needed, help to improve the utility of the evaluation.

2.2 Identify Needed Evaluation Information and Required Type of Evaluation

Table 2-1 illustrates examples of intended uses for evaluation results, the various kinds of evaluation information that could help inform decisions, and the relevant types of evaluations.

Table 2-1. Types of Information Associated with Different Types of Program Evaluations

Intended Use	Types of Information Needed	Type of Evaluation
Make continuous program adjustments to correct implementation weaknesses	• **Measures by which the efficiency and effectiveness of program implementation processes may be judged** This might include, for example, measures of the effectiveness of specific activities, such as speed of contracting, percent target audience reached, and customer satisfaction; what has worked and what has not worked; and where additional resources could be leveraged.	**Process evaluation**
Communicate program's value to key stakeholders	• **Quantitative and qualitative outcomes that can be attributed to the program's outputs** This refers to information about outcomes that would not have occurred without the influence of the program, sometimes called "net impacts."	**Impact evaluation**
Expand or curtail program investments based on knowing where the largest benefits occur for dollars spent	• **Quantitative and qualitative measures of performance relative to funding** Benefits are usually quantified in dollars, but may also include environmental impact reductions and jobs created, ideally with comparable data on different strategies for reaching the same objectives, or to compare benefits and costs of substitutable strategies.	**Cost-benefit studies / Cost-effectiveness studies**

The intended use determines the type of information needed, which determines the type of evaluation to conduct to obtain that information.

2.3 Align Timelines to Ensure that Evaluation Results are Available when Needed

In order to align the evaluation timeline, a conventional heuristic device is to work backwards from the anticipated end of the evaluation study, following these steps:

- Determine when the information from the evaluation is needed for the intended use. For example, is it needed for the project annual operating plan (AOP), for multi-year program planning, or for budget defense?
- Is it needed in six months, 12 months, even 18 months, or as soon as feasible? This time of need, combined with the importance of the use to which the evaluation results would be put, should determine the type of study to be done and the time required to do it optimally (or available to do it minimally).
- Allow time for quality assurance review of the evaluation plan and draft evaluation report (see Steps 3 and 6). Each review can take anywhere from 2.5 to 4 weeks.
- Estimate the time it will take to perform the evaluation. For example, if the evaluation is likely to require a survey to collect data from more than nine non-Federal entities, allow time

The timeline referred to here is the timeline for the entire evaluation process, from determination of the objectives to making the decisions that will be based on the evaluation results (Step 1 through Step 7). The timeline for performing the evaluation itself (Steps 4 6) is part of this overall timeline.

for OMB to approve the survey.[8] OMB approvals have been known to take as much as 6-12 months to secure. Consideration must also be given to the time needed to secure program data. Some program data have taken 2-4 months to secure. Step 4 (Section 5) and Appendix D contain guidance on obtaining OMB clearance to conduct a survey.

- Determine when the evaluation must begin in order to deliver its information when it is needed.
- Account for the administrative time required to hire an evaluation expert, a process that could take 1-3 months.

2.4 Determine the Level of Evaluation Rigor Needed

Evaluation rigor, as used in this Guide, refers to the level of expected reliability of the assessment. It is a measure of whether an assessment is of good quality and findings can be trusted. The higher the rigor, the more confident one is that the results of the evaluation are reliable. Since evaluations must be conducted to suit specific uses, it stands to reason that the most important decisions should be supported by studies whose results will have the highest rigor. EERE has developed a quality assurance rating system for assigning evaluation studies into "tiers of evidence" based on level of rigor.[9] For example, a well executed randomized controlled trial, or an excellently executed quasi-experiment with exemplary treatment of internal validity threats, would be rated as Tier 1 studies.

Criteria for Rating the Level of Rigor of EERE Evaluation Studies

The criteria for classifying impact evaluation studies into levels of rigor includes:

1) The research design (randomized controlled trials [RCTs], quasi-experiments, non-experiments with and without counterfactual)
2) The identification and treatment of internal and external (where applicable) threats to the validity of the study
3) The actual execution of the study in terms of implementation of sampling protocols, data collection and analysis, and quality assurance
4) Any additional steps taken to strengthen the results (e.g., through the use of mixed methods to support the primary design).

Surveys of Federal Government employees about Federal Government activities do not require OMB clearance.

The tiers of evidence are defined as follows:

Tier 1 = Very strong level of rigor. High scientific quality, excellent treatment of internal validity threats, and excellent execution. The equivalent of a well-executed RCT;

Tier 2 = Strong level of rigor. High or moderate scientific quality, with good or excellent treatment of internal validity threats, and good to excellent execution;

Tier 3 = Moderate level of rigor. Intermediate scientific quality, with adequate-to-good treatment of threats to internal validity, and adequate-to-good execution;

Tier 4 = Low level of rigor. Poorly executed evaluation with high, moderate or intermediate scientific quality, and with adequate treatment of internal validity threats, or poorly designed evaluation of limited scientific quality, with adequate execution;

Tier 5 = Very low level of rigor. High, moderate, or intermediate scientific quality with very poor treatment of validity threats, and very poor execution, or a study with very limited scientific quality, and severe vulnerability to internal validity threats.

Source: *Rating the Level of Rigor of EERE Evaluation Studies*. Prepared by Yaw Agyeman (LBNL) for DOE/EERE, August 2015.

An example of appropriate use, based on evaluation rigor, would be to use a Tier 1 study to support decisions involving the highest program priority or most expensive program investments. If a key stakeholder, such as the U.S. Congress or the White House, asks to know the impact of a program investment, the evidence would need to be very strong or strong. In such a case, a Tier 1 or a Tier 2 evaluation study would be appropriate, but not a Tier 4 or Tier 5 evaluation study. Conversely, if the evaluation is to support a decision involving a lesser level of investment or process efficiency, or if the result is expressed only as an outcome (not impact), then a Tier 4 or Tier 3 study might suffice.

2.5 Formulate Initial Logic Model, Metrics, and Evaluation Questions

A program logic model helps facilitate an understanding of the processes by which program activities are supposed to lead to certain outputs and to desired outcomes. The program logic, in addition to the understanding of intended uses of the evaluation and kinds of information needed for the uses, informs the statement of work (SOW) development process.

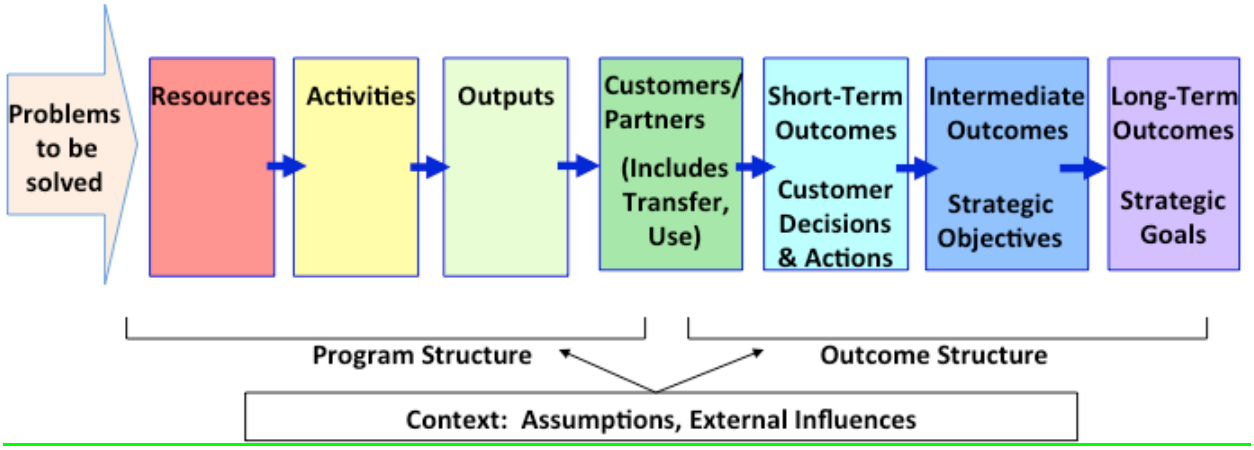
A program logic model is usually a simple diagram (with accompanying text) that identifies the key logical (causal) relationships among program elements and the problem to be solved (the program's objective), thus defining pathways to success. This pathway represents the program's underlying theory of cause and effect. That is, it describes the inputs (resources), activities, and outputs, the customers reached, and the associated sequence of outcomes that are solutions to the problem. The logic also includes factors external to the program that drive or restrain program success.[10]

Construction of a logic model is highly recommended, even in nascent, preliminary form because it makes explicit the relationships between program's activities and its desired outcomes. These relationships help the manager and evaluator identify key metrics and research questions that guide evaluation efforts and lead to an understanding of the outcome results. This initial logic model will also help guide the preparation of the study's statement of work for eventual use in drafting the RFP.[11] Figure 2-1 illustrates the basic elements of a program logic model.

[10] McLaughlin, John A and Gretchen B. Jordan. 2010. "Using Logic Models." *Handbook of Practical Program Evaluation,* 3rd Edition, Wholey, J., Hatry, H., and Newcomer, K., Eds., Jossey Bass, 55-80.

[11] A useful discussion of logic models, including a stage-by-stage process for constructing them, can be found in the W.K. Kellogg Foundation. "Logic Model Development Guide." (2004). Battle Creek: W.K. Kellogg Foundation. Available at: http://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide. Last accessed 4/28/14. The University of Wisconsin–Extension Website also has useful resources on the development of logic models. Available at: www.uwex.edu/ces/pdande/evaluation/evallogicmodel.html.

Figure 2-1. The Basic Elements of a Logic Model



Source: Gretchen Jordan, EERE Program Evaluation Training, 2014

It is conventional practice that during the development of an evaluation plan by the hired independent outside evaluator (Step 3), a complete program logic model is formulated to further guide metric development and refine the evaluation's research questions. The program logic model prepared by the evaluator is often more complete and detailed than the initial one prepared by the DOE evaluation project manager in this Step 1.

2.6 Estimate Evaluation Cost and Other Resources Needed

Evaluation planning requires an estimate of how much a program evaluation will cost. It is good practice to have this consideration woven into each element of the preparation steps. As noted, the intended uses of the evaluation should be the first consideration in preparing for an evaluation. But often there are multiple needs for any program at a given time (potentially multiple uses for evaluative information) all on a limited budget. This also links back to the need to prioritize among the many information needs of the program.

> It may be necessary to revisit this sub step during the design of the evaluation because resources affect the choice of evaluation method. In any event, the evaluation design process must begin with a sense of the resources available.

A key to greater efficiency through this step is to have a long-term evaluation strategy. This can help the program prioritize not only on what evaluations to conduct, but also how to sequence them in relation to multi-year resource expectations.

The cost of an evaluation study depends on several factors, including the intended uses for the results, the level of desired rigor, the availability of data, the scope of the questions for the evaluation, and the scale of the intervention to be evaluated. Although there is no simple rule of thumb for estimating the cost of a given study, some guidelines are provided here to assist the DOE evaluation project manager to arrive at a reasonable estimate of the range of costs for an evaluation. These guidelines are based, in part, on EERE experience and on recommendations from other studies, and involve the simultaneous consideration of:

- The percent of program budget available to spend on program evaluations, for example, as allocated from set-aside funding; and
- The importance of the results that the evaluation will produce.

2.6.1 Cost As Percent of Program Budget

Some state, electric, and gas utility organizations have used a rule of thumb based on percent-of-annual-program-cost, to establish an annual budget for energy-efficiency program evaluations. Sometimes these rules of thumb apply to multiyear program total budgets when a single evaluation will be conducted at the end of the multiyear period. These percentages include all evaluations planned for a year and have ranged from less than 1% to 6% of the total budget for the programs to be evaluated. The average spending on electric EM&V by program administrators in 2011 was 3.6% of total budget for the evaluated programs.[12] The percentages available for state and utility program evaluation budgets suggest that a reasonable spending range for evaluation is 3% to 6% of a portfolio budget.[13] If the evaluation budget were spread across all programs, these percentages would apply as well to specific program budgets. The variation in these percentages reflects many factors, some of which are discussed in this section. A DOE evaluation project manager should view these broad percentages as reasonable ranges for the amount of funds to commit to evaluation activity for a given program or program portfolio.

2.6.2 Cost Factors for Individual Evaluation Studies

Within the limits imposed by the portfolio budget, the factors that contribute to the cost of an evaluation may be grouped into the following categories, which are discussed in turn:

- The type of evaluation (described in Section 1);
- The degree of rigor required for the evaluation results (described in Section 2.4);
- The scope of data-collection requirements (e.g., number of questions, the size of the sample(s) or census (data collection from the entire population of interest), the Paperwork Reduction Act process, and the extent of difficulty of interviewing the relevant population(s) (discussed under Sections 4 and 5, Steps 3 and 4); and
- The analysis and reporting needs.

2.6.3 Cost Variation by Various Factors

Type of Evaluation. Of the three types of evaluations addressed by this Guide – process, impact, and cost-benefit – the most expensive usually will be an impact evaluation. These types of evaluations are the most challenging to perform because of their scope and because they require that estimates be developed of what would have occurred had no program existed. This estimate is determined by experimental or quasi-experimental design, or, failing that, by

[12] State and Local Energy Efficiency Action Network. 2012. *Energy Efficiency Program Impact Evaluation Guide*. Prepared by Steven R. Schiller, Schiller Consulting, Inc., page 7-16. www.seeaction.energy.gov. (Last accessed May 18, 2015.)

[13] *Ibid,* page 7-14. These percentages are consistent with the percentages identified through a review of regulatory findings and reported in the National Renewable Energy Laboratory's *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* Prepared by Tina Jayaweera & Hossein Haeri, The Cadmus Group, Inc. Subcontract report: NREL/SR-7A30-53827, April 2013, page 1-8. http://energy.gov/oe/downloads/uniform-methods-project-methods-determining-energy-efficiency-savings-specific-measures (Last accessed August 20, 2015.)

developing a so-called "counterfactual". One approach to determining a counterfactual is to interview the participants themselves to find out what they would have done absent the intervention. This may be combined with a demonstration of a chronology of events of what the program did at various stages along a logical pathway to outcomes, as well as what change other programs and/or policies influenced on that same timeline.

Defensibility of the Evaluation Results. All EERE evaluations should be able to withstand the criticism of expert peer reviewers. As described in Section 2.4, the ability of an evaluation's results to withstand criticism is based on its rigor. The degree of rigor required depends on whether results are to be used for a major decision about the program. The need for greater defensibility of study results will impose a requirement for greater rigor in the methods used to generate the results. Greater rigor, in turn, will almost always require more resources for data collection, quantitative analysis, and reporting.

Scope of the Information Collection Requirement. An independent third-party evaluator's cost for collecting data for an evaluation will consist of the following data-collection cost factors:[14]

- Accessibility, amount, and quality of existing data, such as contact information, program reports, and output attainment
- Determining which populations need to be surveyed or interviewed
- Developing the research questions and corresponding data requirements
- The degree of precision and accuracy sought for the data measurements which, in turn, influence the sample sizes for each survey (these concepts are described in Section 4.4)
- Satisfying the Paperwork Reduction Act requirements for the Office of Management and Budget (OMB) if the sample will be larger than nine persons
- Obtaining and preparing the sample(s)
- Conducting the information collection(s)
- Preparing the collected information for analysis.

The prices for these components will correlate with the number of variables that must be measured to answer the evaluation's research questions, the difficulty in making acceptable measurements, and the defensibility required for the evaluation results.

A survey of known program participants might expect 50% to 70% of the participants to complete an interview, but when no list of program participants exists, or when a comparison group is being interviewed, the percentage of attempted interviews that result in a completed interview can be quite low. If an impact evaluation also requires a parallel survey of non-participants for comparison purposes that survey might expect 1%-5% of the attempted eligible non-participating population to complete the interview.

Any evaluation that requires collecting the same information from more than nine respondents must be approved by OMB under the requirements of the Paperwork Reduction Act (PRA). This process imposes additional costs on the study. Appendix D provides a more detailed description of the PRA processes, requirements, and points of contact for each.

[14] This Guide follows the practice of the Office of Management and Budget and uses the terms "data collection" and "information collection" interchangeably.

If the defensibility of an evaluation result requires physical measurements such as the actual metering of energy usage, the cost of information collection will be many times greater than the cost of data collected by telephone, records review, or in-person interviewing.

Analysis and Reporting Needs. The following features of an evaluation correlate with the evaluation's cost of analysis and reporting:

- The number of information collections
- The number of variables measured by the information collections
- The complexity of the analyses required to produce evaluation results from the measurements
- The use of statistical tests to support the defensibility required for the results
- The design of the report used to communicate the results and explain the research and analytic methodologies (provided in support of the results).

2.6.4 Typical Cost of an Individual Evaluation Study

The variation possible in the cost factors described in the preceding sections creates large ranges in total costs for the different types of evaluation covered by this Guide. Table 2-2 provides illustrative cost ranges for each of these types for a single evaluation. The right-hand column of Table 2-2 lists some of the factors that will affect the actual cost within the ranges.

Table 2-2. Illustrative Costs for an Individual Evaluation Study

Type of Evaluation with Illustrative Scope	Cost Range*		Other Factors Influencing Cost within the Ranges Shown
	Lower Defensibility	Higher Defensibility	
Process Evaluation Illustrative Scope: customer satisfaction measurement; implementation efficiency	$25,000 - $50,000	$50,000 - $150,000	• Number of populations to be interviewed • Difficulty in identifying and contacting eligible members of the population • Number of questions to be asked • Choice of survey method (e.g., in-person, telephone, mail, Web) • Type of PRA clearance needed
Impact Evaluation Illustrative Scope: quantification of 5-8 direct and indirect outcomes *attributable* to program (also referred to as "net impacts")	$150,000 - $300,000	$250,000 - $600,000	• Number and complexity of outcomes (scope) • The geographic scope of the program's impacts being estimated; a large geographic scope usually will increase the cost of sampling and data collection • Difficulty in completing interviews with the target population(s) • Sources of information (e.g., participant and non-participant surveys) • Availability of a program-implementation baseline • Research design used to control for outside influences (e.g., experimental vs. non-experimental research design) • Method used to estimate net outcomes • Full PRA approval process for surveys • The number of questions asked • The number of different populations to be interviewed • The sampling precision sought
Cost-benefit Evaluation	$75,000-	$150,000 -	• A specific kind of impact evaluation to

Illustrative Scope: Comparison of quantified of energy and environmental benefits relative to associated costs	$150,000	$400,000	quantify the gross or net energy savings or other outcomes • Effort needed to quantify other non-energy benefits (e.g., job creation, environmental emissions reductions) • Ease of modeling or otherwise estimating the costs of the program that produced the benefits • Type of cost-benefit test used,(e.g., societal costs and benefits or participant costs and benefits)

* The cost ranges shown reflect EERE experience over the past five years. However, neither the low nor the high bounds should be considered binding.

Table 2-2 shows the range of costs typical of the three types of program evaluations covered by this Guide. Table 2-3 provides evidence from evaluation studies conducted for EERE, of how typical evaluation costs might be distributed across evaluation tasks. The table shows the average proportions of an evaluation budget devoted to each of eight typical evaluation tasks. The proportions are based on a sample of EERE evaluations initiated between 2008 and 2015. Table 2-3 presents these proportions as average percentages of total labor hours and costs committed to each of the evaluation tasks. The evaluation projects represent a wide range of scope and complexity. To indicate this, Table 2-3 also shows the range of percentages from the evaluations.

Table 2-3. Illustrative Allocation of Costs by Task for EERE Impact Evaluations[15]

Task	Labor Hours as a Percent of Total Labor hours	Task Costs as a Percent of Total Costs
1. Conduct a project initiation meeting with DOE staff to discuss proposed work and schedule	Average: 1% Range: 0.4%–2%	Average: 1% Range: 0.4%–2%
2. Conduct a preliminary review of key documents and hold meetings and interviews with program managers and key stakeholders	Average: 8% Range: 1%–24%	Average: 8% Range: 1%–27%
3. Create draft and final evaluation plan	Average: 14% Range: 7%–35%	Average: 12% Range: 5%–30%
4. Conduct data collection and analysis and provide interim feedback	Average: 44% Range: 3%–60%	Average: 41% Range: 4%–60%
5. Prepare draft and final reports, participate in peer review process	Average: 20% Range: 14%–34%	Average: 22% Range: 12%–29%
6. Prepare summary presentation and brief DOE	Average: 3% Range: 1%–9%	Average: 4% Range: 2%–10%
7. Manage the project	Average: 5% Range: 2%–7%	Average: 7% Range: 2%–22%
8. Provide regular project status reporting	Average: 4% Range: 1%–7%	Average: 5% Range: 1%–13%
Totals	100%	100%

[15] Labor hours are presented for 10 evaluation studies, while task costs are presented for 22 studies. Average travel cost for 17of the studies (usually for purposes of meeting stakeholders in DOE/EERE) was 2% of total costs, ranging from 0.2%-3%.

The *labor* percentages in Table 2-3 exclude any major non-labor costs. Evaluators often subcontract data collection to vendors that specialize in data collection. When this happens, data collection may add 27% of the labor cost to the total project cost.

2.7 Organize Background Data and Program Records

One of the costliest aspects of conducting an evaluation study is the acquisition of valid, complete, and quality-assured data to answer questions the study is designed to answer. The costs arise from the convergence of several difficult tasks:

- Routinely collecting basic data in a standardized format
- Obtaining a large enough sample to provide sufficient precision and statistical power for the measurements and hypotheses of interest[16]
- Overcoming non-response and recall bias from participants and non-participants
- Undertaking ad hoc efforts to assure data quality.

Some of the cost may be attenuated if provisions are made for routinely gathering key information from the study's participants during program operations. The cost of constructing an ad hoc database of the program outputs and outcome history at the time of the evaluation can be significant. If program outputs and outcome data have been collected and recorded in a useable database from the beginning of the program, the cost of an evaluation may be reduced significantly (and the ease of real-time program performance monitoring will be increased).

It is in this interest that EERE is now actively including evaluation information in the new central information system. Programs are encouraged to participate in the development and maintenance of the data (metrics and associated measures) to be routinely gathered for both performance monitoring and for use in current and future evaluation studies.

[16] As a general convention, the degree of confidence used is 95 percent, with 80 percent power.

3.0 STEP 2. HIRE AN INDEPENDENT OUTSIDE EVALUATOR

This section recommends a process for hiring an independent, outside evaluator to perform an evaluation study. Briefly, this involves using a Request for Proposal (RFP) process to select a qualified independent third-party evaluator, typically through a competitive solicitation process.

RFP's generally include the following elements:

- Program background
- Objective of the RFP
- Statement of Work
- Basis for selection / evaluation of proposals
- Request for references
- Proposal format and other preparation instructions
- When and where to submit proposals.

This is also the appropriate time to ensure that a procedure for external peer review is created for the evaluation (see Section 3.3). The guidance provided by this section covers the technical portions of the RFP.[17]

3.1 Implement Competitive Solicitation Process to Hire an Evaluator

Independent external expert evaluators usually are hired through a competitive solicitation process. In some rare instances, particularly when the resources for the study are limited, a sole source contract might be used instead to identify an expert with no conflict of interest whose considerable expertise means that the learning curve for conducting the study would be minimal, thus optimizing the scarce resource towards the objectives of the study.

The process begins with the development of an RFP (see Section 3.2), which is broadcast either to the entire evaluation community or to a limited number of experts who are expected to have the requisite qualifications.[18] Concurrently, the evaluation project manager selects a team of 3 to 8 experts representing the right balance of pertinent knowledge (subject matter experts, evaluation experts, statisticians, etc.) to serve as reviewers.

There are at least two rounds to the RFP review process. First, each expert reviews all the responses and submits their ordered ranking of the proposals, from strongest to weakest. In a subsequent live debate, the reviewers provide justifications for their views on the proposals. This round ends in a winnowing down of the proposals to the consensus top two or three.

Secondly, since all proposals ultimately have some weaknesses, those making the cut are asked to address aspects of their proposal that were deemed to be weakest. They do this usually

[17] This section does not cover the DOE procurement process (except the when and where to submit proposals) or the terms and conditions of DOE contracts. If the evaluation will be competitively sourced to an independent third party evaluator through DOE's procurement process, the program manager should work with DOE's procurement and contracts offices to ensure that DOE's procurement procedures are followed and that the RFP includes DOE's terms and conditions.

[18] A request for qualifications may be issued to the entire evaluation community beforehand to help determine which experts are likely to have the requisite qualifications and interest.

through both a written response and an oral presentation, depending on the importance of the evaluation, presenting cost-effective and potentially innovative solutions to the areas of concern that were highlighted. This represents the second round of review. After the second round of review, the team of expert reviewers meets again to debate the merits of the revised proposals and to vote for the proposal they believe most persuasively addresses the reviewers' critiques. Then, the chosen independent third-party evaluator is hired in accordance with DOE's procurement regulations.

3.2 Develop the Request for Proposal (RFP)

The following are some of the details typically found in an RFP:

- **The program's background**. This covers the history, mission, goals, and objectives of the program to provide the proper context for the evaluation.
- **The objectives of the evaluation**. The objectives describe the broad uses prompting the need for an evaluation and its goals, defined in such a way as to be measurable. The list of objectives defines for the independent third-party evaluator the purposes that the program manager wants the evaluation to serve and, therefore, constitutes a critical piece of information governing the evaluation project.
- **The Statement of Work (SOW)**. The SOW outlines the scope of the evaluation and describes its specific requirements. It often specifies the tasks expected for performing the evaluation. A common set of tasks will help the proposal reviewers compare proposers' understanding of the evaluation's components and their capabilities for performing them. The SOW might be revised during discussions between the DOE evaluation project manager and the successful evaluator. Example SOW's are shown in Appendices A-1 and A-2. The following constitute some of the SOW elements that will help the bidders prepare responsive proposals:
  - ***Initial evaluation metrics.*** The objectives of an evaluation and program logic suggest key metrics of desired results to measure and calculate. The program manager may suggest evaluation metrics to satisfy the objectives, but expect the evaluator to propose other metrics as well.
  - ***The evaluation questions and their priorities.*** Specific questions for the evaluation flow from the evaluation objectives and program logic. An example of a process evaluation question might be "What is the efficiency of getting grant funds out?" An impact evaluation question example might be, "Did these outputs cause the observed outcomes?" For impact evaluations, the questions should relate to the types of direct and indirect outcomes to be evaluated (based on program theory/logic model). The evaluator may restate the questions in forms that allow for more accurate measurement (i.e., as detailed research questions).
  - ***An evaluation plan***. The independent third-party evaluator must develop a full evaluation plan (Section 4, Step 3) incorporating key metrics and questions and methodologies. Whenever possible, relevant lessons learned from previous program evaluations should be incorporated into the section of the RFP requiring the evaluation plan.
  - ***Alternative, complementary, innovative methodological approaches.*** Some evaluation questions might have obvious, validated methodological approaches for answering them. However, it is always advisable to invite creative, alternative and particularly complementary methodological approaches to strengthen the certainty of the findings.

- ***Reports and other deliverables required***. This includes periodic performance and budget reporting. One of the deliverables must be the evaluation plan (Step 3).
- ***Resources that the EERE evaluation project manager will provide to the independent third-party evaluator*.** Examples include: participant lists; records of outputs and outcomes; expenditure records; and access to program staff for interviews. Having such resources available informs bidders on the scope of data collection required and therefore on estimated costs.
- ***The EERE Quality Assurance (QA) Plan*.** The SOW should require the independent third-party evaluator to develop a QA plan, but the evaluation project manager should also have one that includes peer reviews of the draft evaluation plan and study report, in conformance with established EERE guidance for conducting and reviewing evaluation studies.
- ***Initial evaluation schedule and milestones*.** Include a milestone for the kickoff meeting with the independent third-party evaluator to discuss the above topics. The due date for the final report should take into consideration the date of any decision whose outcome may benefit from the evaluation's results. A presentation to stakeholders after the final report may be useful. Build into the schedule the time required for quality assurance, including for reviews of the evaluation plan and the draft final report.

- **Potential technical challenges or problems that may be encountered for the type of evaluation requested, and bidders proposed resolutions for these**. Recognition of potential problems or challenges and resolutions will illustrate the bidder's experience levels and capabilities to address study issues as they arise, and help them plan the evaluation. Examples might include collecting data from states or from non-participants; dealing with issues that arise when billing data are used; a design that will permit estimation of attribution (for impact evaluations) with the desired level of rigor; designing a probability sample; use of savings ratios; and dealing with potential survey non-response issues.
- **Evaluation criteria.** The evaluation project manager should specify the criteria on which proposals will be judged and may include a point system for weighting each criterion. This will help produce comparable proposals and give the proposal reviewers a set of common criteria on which to base their judgments. DOE's procurement office may also contribute requirements to the evaluation criteria.
- **List of references.** Usually the evaluation project manager will require that the bidder provide a list of two to four references to managers of other evaluation contracts that the bidder has performed. This requirement may specify that the reference contracts be within a recent time period.

Program managers sometimes ask bidders to provide examples of evaluation reports to help them assess the ability of the bidder's organization to write clear reports. This may reduce the number of bidders, however, as such reports are often proprietary.

- **Proposal format and other preparation instructions.** This feature of an RFP tells the bidders how the program manager requires that the proposal be organized. Such instructions may provide another common basis on which to judge competing proposals. For example, this is where the RFP may require the following:
  - Organization by specified tasks, if any
  - A page limit on the bidder's proposal

- Specific fonts and spacing
- Placement of specific features in separate sections and the order of these sections
- DOE's contracts and procurement offices may also specify preparation instructions to help them evaluate compliance with the proposal requirements of their offices.

- Where and when to submit proposals. The procurement office will set these requirements in conjunction with the project manager's timetable.

The following additional requirements and information might be included if the DOE evaluation project manager wants to specify greater detail about the evaluation's requirements:

- **Consistency in the use of terminology and between requirements.** If the RFP uses technical terms that a bidder may misinterpret, a glossary will help to reduce misunderstandings and the number of follow-on questions from prospective bidders.
- **Price.** The proposal manager may wish to specify the maximum budget for the evaluation contract. This will also help reviewers compare the proposals on a common base. If low price will be a heavily weighted criterion, that should be mentioned in the evaluation criterion.
- **Types of information required when answering individual specific questions.** Examples of such information include counts, averages, and proportions.
- **Required level of statistical precision for survey results.**
- **Required tests of significance for statistical relationships.**
- **Data-collection and analysis methodologies.** If the project manager expects the independent third-party evaluator to use specific methodologies to answer certain evaluation questions, the methodologies should be specified. Such a specification might occur if Tier 1 or 2 levels of rigor is required. Usually, however, the evaluation manager will rely on the bidders to propose appropriate methodologies.
- **Relevant guidance or references that will give the evaluation expert information about the requirements of Federal program evaluations**. For example, if the evaluation will need to comply with OMB or congressional requirements, provide prospective bidders with the web link(s) to the documents specifying the requirements.

Sometimes independent third-party evaluator support is needed after the final report is accepted. The DOE evaluation project manager may ask the evaluation bidders to propose separate time and materials rates to provide support related to the evaluation after the project is over. However, such support should never involve correcting technical or factual errors in the evaluation. Any and all such errors are to be addressed by the third-party evaluator over the course of the study implementation and quality assurance review.

3.3 Ensure EERE Quality Assurance Protocol is Set Up for Implementation

This step – an activity for the DOE project manager sponsoring an evaluation study – is essential to ensure that the evaluation results are defensible, with consideration given to the resources that are available for it. The EERE Quality Assurance Protocol specifies how the data collection, analysis, and reporting activities will themselves be peer reviewed by external experts who are not part of the evaluation team.

Although establishing a quality assurance protocol for the study is not directly related to hiring the third-party evaluator, it is best to do so concurrently, to ensure that there is adequate time to identify the best reviewers for the study, as part of establishing the best protocol.

A well defined quality review process must be in place before the evaluation begins.

For the DOE project manager sponsoring[19] an evaluation study, the following quality assurance (QA) guidance applies:

- Use independent third-party evaluators who are objective, with no real or perceived conflict of interest (COI). Independent third-party evaluators who have a long-standing relationship with an EERE program that includes involvement in daily or routine program implementation and analysis activities generally would not be considered independent third-party evaluators without special exception. If allowed to bid for an evaluation, such independent third-party evaluators should be asked to sign a COI form.
- Independent third-party evaluators are expected to prepare a detailed evaluation plan (Step 3, Section 4), and participate in a peer review of the draft evaluation plan and draft evaluation report. The peer reviewers selected for the evaluation should be assembled to fully scrutinize the independent third-party evaluator's evaluation plan, execution, and reporting.

DOE has two options for constituting peer review panels.

- Establish a ***standing*** peer review panel. This panel may comprise broadly experienced evaluation experts who are "on call" to act as peer reviewers for the evaluation plans and final reports of several evaluations or for either part of an evaluation.
- Identify an ***ad hoc*** panel of three to eight specially selected external evaluation experts to review and provide written comments on the draft evaluation plan and/or the draft evaluation report for a single evaluation. Such individuals might also be experts in the technology whose development is the objective in a deployment program, in which case they could be chosen to complement a standing review panel.

The evaluation project manager may also select a team of internal stakeholders (e.g., program staff and/or national lab experts associated with the program) to serve as internal peer reviewers. These reviewers will not be independent, but their special knowledge may point out ways to improve the product.

The objectivity of the process can be aided by creating a list of specific "criteria" that the reviewers must address for both the evaluation plan and the draft report. Minimum criteria include:

Research Design
Key requirements are ensuring that the methods and procedures employed to conduct the evaluation study are appropriate. Inherent to this is the requirement that the research questions are well formulated and relevant to the objectives of the evaluation, and that the metrics are credible as measures of the outputs and outcomes required to satisfy the evaluation's objectives.

[19] "Sponsoring" means the EERE program provides the funds for a study and has a staff that has responsibility for managing the contract of an independent outside evaluation professional. The evaluation professional conducts the study. It is not an option for program evaluation studies to be conducted only internally by EERE staff.

For statistical methods, the degree of relationship between indicators, tests of significance, and confidence intervals (statistical precision) for sample estimates, should be built into the analysis and applied wherever possible. The evaluation plan must demonstrate understanding of previous related studies, and the data collection and analysis methods must be credible.

Treatment of Threats to Validity

The threats to the internal validity of a study refer to the various sources of bias that might undermine the validity of claims made in the evaluation, including claims of attribution. In effect, a study that fails to identify and remedy the potential threats to its internal validity cannot be deemed to have validly and reliably asserted that the conclusions about the process or outcomes are true. Key among these threats are:

- Temporal antecedence (effect does not precede cause);
- Selection bias (effect is not due to systemic differences between participants and non-participants); and
- Confounding (all other known rival explanatory factors are controlled for).

Other internal validity threats such as history, testing, contamination, differential attrition, regression-to-the-mean, instrumentation, "John Henry effect," resentful demoralization, selection-maturation interaction and selection-history interaction, can also adversely affect whether the findings of a study are valid.[20]

Additionally, evaluation studies for which the results from the study population are intended to be generalizable across other populations, settings and timeframes must appropriately address the threats to external validity. Examples of threats to external validity include the interactive effect of testing, the interactive effects of selection and treatment, and multiple treatment interference.[21] Failure of the study to address these threats would make the findings, even if they are internally valid, unsuitable for generalization across other populations, settings, and time.

Execution

Quality assurance also covers the execution of the evaluation study. Execution refers to the actual use of the planned protocols for implementing the evaluation, namely data collection protocols, measurement methods, analysis approaches, and reporting of the results, including the conclusions drawn on the basis of the analysis. These criteria—data collection approaches, measurement methods, and analytical approach—are subject to critique during the review of the evaluation's plan. The methods and approaches should have been implemented during the study unless departures from them are explained in the draft report and the departures can be judged reasonable. The following exemplify these criteria:

- Data Collection
    - Were all planned data collected as proposed? If some values are missing, how they were treated?
    - If missing data values were inferred, was the inference method appropriate?

[20] Shadish, William R., Cook, Thomas D., and Donald T. Campbell. 2001. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Cengage Learning; 2nd Edition.
[21] Ibid

- Was the data inspected for out-of-range values (outliers) and other anomalies, and how they were treated?
- How was non-response addressed, if it was an important issue for the study?
- Were the data collection methods actually implemented as planned, or if revisions were required, were they appropriate and the reasons for the revisions documented?
- Were all collected data provided and their layout documented?

- Analysis
  - Were the analysis methods actually implemented as planned, or if revisions were required, were they appropriate and the reasons for the revisions documented?
  - Was the documentation of the analytical approach accurate, understandable, and reasonable?

Reporting Criteria

Quality assurance also includes ensuring the quality of the report, and covers the following:

- Is the evaluation plan and draft report easy to read and follow?
- Is the report outline draft appropriate and likely to present the study findings and recommendations well, and to provide documentation of methods used?
- Are the calculations and data presented in Tables fully documented and transparent?
- Do the draft findings and recommendations in the evaluation report follow logically from the research results and are they explained thoroughly?
- Does the draft report present answers to all of the questions asked in the evaluation plan, as revised through the work plan?

Consideration of all of the quality assurance criteria listed above during the review of the evaluation plan and draft report provides the basis for classifying evaluations into the tiers of evidence (1-5, highest to lowest) corresponding to their rigor, and supports the overall confidence in the evidence they provide in support of the evaluation's objectives. These tiers of evidence, in turn, enable managers to put the evaluation results to the uses for which they were intended, for either program improvement or accountability.

The review steps where these QA criteria will be examined should be included in the evaluation plan developed under Section 4, Step 3. These quality assurance protocols are indispensable to the goal of obtaining a useful and defensible evaluation product.

4.0 STEP 3. DEVELOP AN EVALUATION PLAN

This section provides guidance on the development of an evaluation plan, covering the essential elements that go into the plan. This step is the responsibility of the independent third-party evaluator, but the DOE project manager is advised to become familiar with elements involved in developing an evaluation plan. These elements include a more detailed logic model, the development of metrics from the logic model, and the formulation of specific researchable evaluation questions. Once the evaluation research questions have been formulated, the next challenge is determining an appropriate research design for the study, a data collection plan, and an approach for analyzing the data. The draft evaluation plan is then subjected to the peer review process described in Section 3.3.

Elements of the evaluation plan described in this section include the following:

- Develop a final program logic model, metrics, and researchable evaluation questions
- Perform an evaluability assessment
- Determine an appropriate evaluation research design
- Establish a data collection plan
- Choose the appropriate analytical method(s) for the selected research design
- Participate in an external review of the evaluation plan.

4.1 Develop Final Logic Model, Metrics, and Researchable Questions

At this stage in the project, the independent evaluator has been hired. The evaluator's task begins with gathering program records, engaging with the manager of the program and possibly with other program stakeholders, and preparing the final logic model. As mentioned in Section 3, this final logic model will typically be more detailed and refined than the initial logic model developed by the DOE evaluation project manager. The more detailed logic model will facilitate the identification of metrics and be used to refine the initially formulated evaluation questions. This brief encapsulation covers what the evaluator would typically do in preparing the final logic model:

- Gather program records and other documents, engaging with the manager of the program and possibly with other program stakeholders
- Prepare the final logic model at an appropriate level of detail
- Identify impact and/or process metrics (depending on study scope), including revisiting and possibly refining the metrics created earlier by the DOE evaluation project manager in Step 2 (Section 3.1)
- Formulate high-level evaluation questions for the study, and prioritize them (revisiting and possibly refining the questions created earlier by the DOE evaluation project manager in Step 2)
- Prepare specific, researchable questions the evaluation must answer through its data collection and analysis.

Figure 4-1 presents an example of a logic model for the EERE's Better Buildings Neighborhood Program (BBNP). The logic model is offered from the grantee's perspective, identifying the set of activities that the various funded grantees undertook, along with the expected outputs and outcomes (short-term, intermediate and long-term). Metrics for the outputs and outcomes emerge

from the program logic and suggest researchable questions that will ultimately permit the independent third-party evaluator to satisfy the evaluation's objectives.

Developing researchable questions (i.e., the specific framing of the evaluation metrics into specific questions that can be tested) must be addressed next. The researchable questions should be aligned with the metrics identified as needed to satisfy the evaluation's objectives. As an example from a different EERE program, Table 4-1 presents examples of research questions and associated metrics (some of which are derived from other metrics, such as wind power additions since base year) evaluated for EERE's Wind Powering America (WPA) initiative.

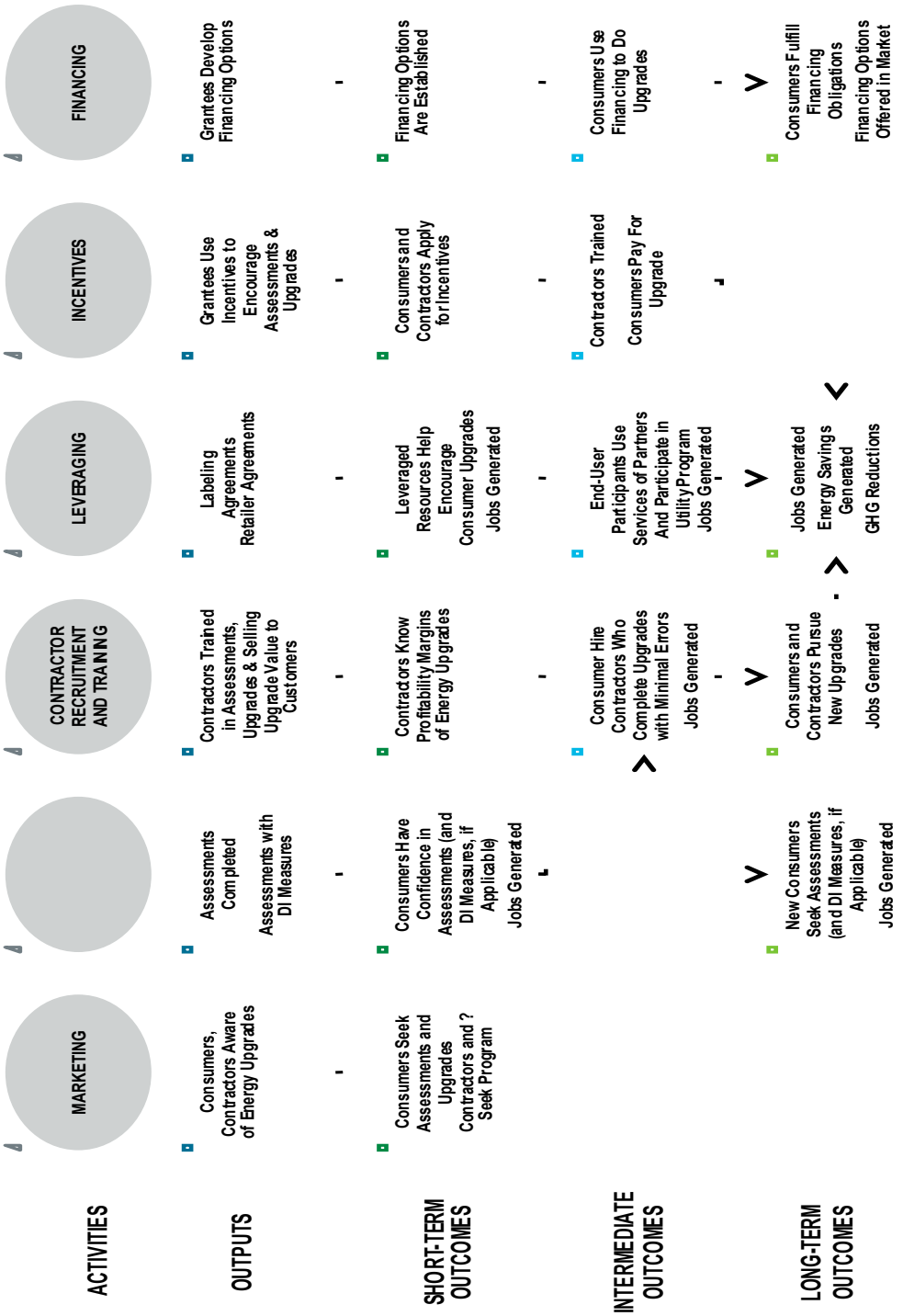
Table 4-1. Examples of Metrics and Associated Research Questions

Research Questions	Metrics Evaluated
What has been the megawatt (MW) capacity growth in states that were influenced by WPA activities? Was a portion of the influence from other market factors (e.g., a state's adoption of a renewable portfolio standard (RPS) related to WPA's influence?	• Percentage-based share and capacity-equivalent estimate of wind power additions influenced by WPA state-based activities and wind working groups (WWGs) according to interviewed stakeholders • Stakeholder estimates of how many fewer MWs would have occurred in a state (or how much later they would have occurred) had WPA and the WWG not existed
What is the perceived level and importance of resources or dollars leveraged by the States from DOE's investment for wind energy deployment activities?	• Stakeholder Likert-scale* ranking of the importance of third-party funds and resources toward the success of a WWG's activities • Stakeholder estimates of how many fewer MWs would have occurred in a state (or how much later they would have occurred) had the WWG not secured additional resources and funding
What is the extent of replication that has occurred?	• Number of organizations or partnerships formed with similar objectives as WPA and WWGs, including those in non-targeted states • Number of organizations that have adopted WWG activities or tactics following reduction or elimination of DOE funding
What elements of WPA's state-based activities have been most successful and why?	• Frequency of unaided recall of WPA activities by interviewed stakeholders • Likert-scale rankings of state-based activity components
Which WWGs have been most successful and why? What are the characteristics of the successful WWGs that fostered their effectiveness?	• Comparison of WPA's share of influence on capacity additions in each state • Open-ended, qualitative responses from respondents in states with high WPA influence
What, if any, common conditions were present for states where the WWGs were less effective? What could be done to minimize these conditions in the future?	• Comparison of WPA's share of influence on capacity additions in each state • Open-ended, qualitative responses from respondents in those states with low WPA influence
*A Likert scale is a tool commonly used in surveys whereby a respondent provides a response to a question using a predefined ranking scale.	

Source: Navigant analysis[22]

[22] "Impact and Process Evaluation of the U.S. Department of Energy's Wind Powering America Initiative." Prepared for DOE by Charlie Bloch, Frank Stern and co-authors, Navigant Consulting, Inc. Washington, DC: 2013.

Figure 4-1. Better Buildings Neighborhood Program High-Level Logic Model from a Grantee Perspective



Source: "Preliminary Process and Market Evaluation: Better Buildings Neighborhood Program." Prepared by: Peters, JS, Bliss R, McRae M, Hoefgen L, Clendenning G, and Barclay D. (2012). Berkeley: Lawrence Berkeley National Laboratory.

4.2 Perform an Evaluability Assessment

Some programs may not be ready for evaluation, despite the desire a program manager might have to commission an evaluation study; others may pose costly challenges to a defensible evaluation. Such situations can exist for a variety of reasons. For example, it may be that the program's goals are not well defined or are unrealistic, or that the program has not been operating long enough to have produced impacts, or that data needed for an evaluation have not been recorded. An evaluability assessment (EA) is a pre-evaluation examination that is intended to assess whether a program evaluation has a reasonable chance of being successfully evaluated and producing defensible and useable results given the resources available for the evaluation.[23] The evaluability assessment will recommend whether the evaluation can proceed as anticipated, whether the research design has to be re-considered, whether the overall evaluation has to be reduced (or increased) in scope, and whether the evaluation's objectives should be modified.

Funding for an EA usually is part of the funding for the evaluation. Evaluability assessments are most useful in the following types of program evaluation situations:

- Uncertainty about availability and quality of data needed to support the analysis
- Existence of many different audiences, necessitating a prioritization of available evaluation resources to meet the needs of the most important audience(s)
- No clear or well-articulated policy goals served by the program
- No well-articulated theoretical connections between the program's activities and the intended outcomes
- Creating operational definitions of the outcome indicators to be measured is expected to be difficult, controversial, or very expensive
- Intended use of the evaluation's results is vague
- The program has multiple stakeholders who may hold differing perceptions of the program's objectives
- Stakeholder interest in the evaluation is weak.

A typical EA consists of the following steps:[24]

- Determine what the EA should accomplish, secure a commitment to conduct it, and establish an ad hoc or formal working group comprised of the evaluation team, the program administration, and the program stakeholders as necessary.
- Define the geographic, population, resources, and other boundaries of the program to be evaluated. This will help to focus both the EA and the subsequent evaluation. This step may result in limiting the evaluation to outcomes within selected boundaries.
- Identify and analyze the program's documents. This will help to clarify the original intent of the program and any subsequent changes that have occurred in its goals. This step will inform agreement on the evaluation's objectives.
- Clarify or develop the program theory. This establishes the assumptions regarding how the program's resources, activities, outputs, and perhaps intermediate outcomes are expected to

[23] In the academic literature, an evaluability assessment often includes such broad assessment goals as to constitute a complete process evaluation. This Guide treats it as an assessment of the feasibility of producing a defensible program evaluation (of the kind being considered) at the time the assessment is performed.

[24] The steps are adapted from Trevison, M.S. and Y.M. Huang. "Evaluability Assessment: A Primer." *Practical Assessment, Research & Evaluation*. 8(20). 2003. http://pareonline.net/getvn.asp?v=8&n=20. The authors attribute the steps to Smith, M.F. *Evaluability Assessment: A Practical Approach.* Clemson: Kluwer Academic. 1989.

interact to achieve the program's ultimate goals. These interactions may be depicted in a logic model.
- Determine data availability and the feasibility of accessing them. Can data be collected that will reasonably represent the program activities and outcomes? Will the data have acceptable quality?
- Identify and interview a small set of initial stakeholders or industry experts. Determine what they know and perceive to be true about the program effects. Let them know an evaluation is planned.
- Draw conclusions and make recommendations regarding the ability to conduct an evaluation that will satisfy its objectives (e.g., for intended use and level of rigor).

Evaluability assessments can be done at varying levels of rigor, depending on the scope and level of investment in the evaluation study itself. Quite often they are done informally, at low-level cost, and are rolled into the background research that the evaluator always does in preparing for an evaluation. There are instances, however, when an evaluability assessment might merit a separate, pre-evaluation study budget and effort.

4.3 Determine an Appropriate Evaluation Research Design

The evaluation research design is a research strategy that allows findings to be deduced from the evaluation data. It encompasses the method and procedures employed to conduct scientific research.

The independent third-party evaluator must select an evaluation design that is compatible with the objectives of the evaluation study as well as desired level of rigor and resources or time constraints. The three types of evaluation research design typically used in EERE evaluations are:

- Experimental designs
- Quasi-experimental designs
- Non-experimental designs, with and without counterfactuals.

Each evaluation design, its guidelines for use, and its relative defensibility, are briefly discussed below.

The evaluation contractor must propose a logical approach to inferring answers to the evaluation questions from the data collected by the evaluation study. This logical approach plus the data collection method(s) and analytical method(s) constitutes the "research design." This step discusses the different forms of research design available for impact evaluations.

4.3.1 Experimental Designs

The most important condition for establishing causality between a program activity and its effect is randomization of the individuals or objects who/that will be exposed to the program's activities and outputs (the treatment group) and a group from the same population who will not be exposed to the program (the control group). Randomized controlled trials (RCTs) are designed purposely to meet these evaluation conditions. Randomization, implemented correctly, allows for the creation of two or more program groups that are similar to each other on all characteristics and experiences other than the fact of exposure to the program's intervention (the treatment). This ensures that any outcome differences observed between the groups at the end of

the evaluation are likely to be due to the treatment and not to differences between the groups or to other factors besides the treatment.

As such, they constitute the strongest of the evaluation research designs and lead to the greatest confidence in the estimated program outcomes. RCTs with properly measured and evaluated metrics can produce an estimate of the size of a program's effect that has desirable statistical properties (estimates of the probability that the true effect falls within a defined confidence interval).[25]

RCTs are used whenever strong confidence in a program's actual effect is highly important to national, departmental, or other similarly major interests. The Office of Management and Budget (OMB), which must address evaluations across all government programs, recommends experimental designs whenever feasible.[26] RCTs are common in the education, health, and agriculture fields but are rarely seen in the energy field.

4.3.2 Quasi-Experimental Designs

When a RCT cannot be used, the independent third-party evaluator must develop an approach that approximates an experimental design.[27] In such cases, group assignments for evaluation purposes are often made to approximate the scientific benefits of randomization. For example, program non-participants (called a "comparison" group for quasi-experimental designs) and program participants may be matched on characteristics believed to be correlated with the observed outcomes. There are a variety of research designs using comparison groups. They may be broadly categorized as:[28]

- **Before-After (Pre-Post) Comparison Group Design:** Compare program-participants and non-participants on pre- and post-treatment measurements. Program participants and comparisons are measured at the same time periods. The program effect is deduced by comparing the performance of the participant group pre- and post-intervention with the comparison group. Regression-discontinuity is the strongest of these variations.
- **After-Only (Post-Only) Comparison Group Design:** A less defensible variant of this design simply compares the two groups at the same point after the participants participated in the program, usually because it was not possible, for a variety of reasons, to get pre-test

[25] Shadish WR, Cook TD and Campbell DT. (2002). "Experimental and Quasi-Experimental Designs for Generalized Causal Inference." Belmont, CA: Wadsworth Cengage Learning.
[26] Office of Management and Budget, "What Constitutes Strong Evidence of a Program's Effectiveness?", p. 1. www.whitehouse.gov/omb/part/2004_program_eval_pdf.
[27] When participants voluntarily participate in a program, a type of bias called "self-selection bias" enters into the results. This bias alludes to the probability that the participants have a predisposition to be interested in the program's intervention (e.g., have a prior interest in energy efficiency). This creates two issues for the validity of the evaluation results: (1) it is expensive to find *non*-participants with similar predispositions for a comparison group, and (2) even if they were identified, the results could not be generalized to the broader population group that the program might be targeting because the part of the population without this predisposition would not be included in either group. When this issue is relevant, the independent third-party evaluator should acknowledge that it is a potential source of unknown bias.
[28] The names for these designs have been adapted from D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally & Co., 1966.

measures. The program effect is deduced by comparing the outcomes from the two groups, but only in the period after the intervention, not before.[29]

4.3.3 Non-Experimental Designs

Research designs that do not use control or comparison groups are considered to be "non-experimental" designs.[30] Non-experimental designs can be implemented with or without a counterfactual (some means of estimating what might have been in the absence of the intervention). In non-experimental evaluation designs with counterfactual, the evaluator seeks to obtain an estimate of what might have occurred in the absence of the intervention through the use of one or more approaches, such as using time series for participants only, interviewing the participants, interviewing independent experts, or constructing a statistical comparison group. A mixed method non-experimental approach applies more than one of these non-experimental designs in the same evaluation, with the aim of bolstering the overall findings through complementary lines of evidence, especially if each method points to the same estimates. The use of results from more than one non-experimental approach to develop evaluation findings adds subjective credibility to the findings. Such use of multiple methods in an evaluation is often called "triangulation."[31]

Although non-experimental designs do not establish causality, given the nature of public sector investments, there are instances where non-experimental designs are the only option for evaluating impacts. If they are properly executed and include a method of estimating a counterfactual outcome, they will provide reasonably valid findings on the contribution made by the program intervention on the outcome. At a minimum, non-experimental evaluation designs must include a counterfactual; EERE strongly discourages the use of non-experimental designs without counterfactual.

In sum, per OMB guidelines, experimental designs are the best type of research design for demonstrating actual program impact.[32] RCTs, however, are not always feasible, and in some cases, they are actually illegal or immoral. When RCTs cannot be used, quasi-experimental designs represent the next best category of research methods, followed by non-experimental methods with counterfactual. Several factors – from the intended uses of the evaluation results to the findings from the evaluability assessment – come together to determine what is the most

---

[29] There are several types of experimental and quasi-experimental designs. Determining which is best for different evaluation findings is beyond the scope of this guide. If you have not had prior training in experimental research design, but believe you need to conduct an impact evaluation, it is recommended that you seek expert assistance in assessing the options, or leave the choice of approach to the evaluation expert(s) who propose(s) the evaluation. A good introduction is found in chapter 3 of GAO's "Designing Evaluations," (GAO-12-208G: Published: Jan 31, 2012). http://www.gao.gov/assets/590/588146.pdf.
A more technical, but understandable and short overview is presented in: Shadish WR, Cook TD and Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference: 2nd Edition. Belmont: Cengage Learning, 2002.

[30] Office of Management and Budget, "What Constitutes Evidence of a Program's Effectiveness?" p. 3. www.whitehouse.gov/omb/part/2004_program_eval.pdf. The second definition of non-experimental design given in the OMB document, "indirect analysis" using an independent panel of experts, is more appropriate for R&D projects.

[31] Greene, J., and C. McClintock. 1985. "Triangulation in Evaluation: Design and Analysis Issues. Evaluation Review, v9, no. 5. (October): 523-45.

[32] Ibid.

appropriate research design. The design chosen by the independent third-party evaluator will influence the data collection options described in the next section.

4.4 Establish a Data Collection Plan

Data collection is the process of taking measurements, primarily on the metrics that will be used to answer the research questions discussed in Section 4.1.[33] It includes any activity that produces information that can be used to answer evaluation questions, e.g., surveys, informal interviews, focus groups, and compilation of data from program records. The figure below presents a visualization of the connection between acquisition, design and data collection.

Stronger Evidence

Assessment Question	Common Designs	Likely Methods
Is the program achieving its desired outcomes or having other important side effects?	Compare program performance to standard or expectations	• Bibliometric analysis • Expert review; case study • Surveys or interviews • Anecdotes, Self reporting
	Assess change before and after the program intervention	• Pre, post bibliometrics • Pre, post, or post only Surveys with statistical analysis
	Assess differences between participants and non participants	• Surveys, interviews
Is the program responsible for (effective in) achieving improvements in desired outcomes?	Compare (change in) outcomes for participants and a comparison group	• Can be done with bibliometrics • Case study • Surveys
	Compare (change in) outcomes for participants before and after the intervention, over multiple points in time with actual or statistical controls	• Economic study using data on key variables and possibly interviews • Impact study

The DOE evaluation project manager may choose to leave most of the decision-making for data collection to an evaluation expert; however, a basic understanding of the commonly used alternatives will help the manager judge the proposals offered.

4.4.1 Sources of Data

Data are referred to as "primary" if they are collected specifically for an evaluation (or other purpose) and "secondary" if they are collected by another project for another purpose but will, nonetheless, be useful for the evaluation.

[33] "Data collection" includes any activity that produces information that can be used to answer evaluation questions, e.g., surveys, informal interviews, focus groups, and compilation of data from program records.

Primary Data

The defensibility of the measurement method will contribute to the overall defensibility of the evaluation results. A variety of methodological options exist for collecting data on (measuring) the indicators. Table 4-2 lists the more common methods, along with several of their key characteristics.

Table 4-2. Comparisons of Data-Collection Methods

Data Collection Method	Relative Defensibility	Relative Cost for a Given Survey Size	Comments
Surveys	Varies	Varies	In this table and Guide, the term "surveys" means the collection of data in accordance with generally accepted methods that support statistical inferences (e.g., random sampling for the selection of respondents). The table uses the term "interviews" for data collection without statistical requirements. Both terms presume the use of data collection instruments designed in accordance with generally accepted principles of valid and reliable data measurement.
In-person surveys	High	Usually high	Confidence in the accuracy of the measurements is usually highest of survey types. In the case of measurements of subjective data such as opinions, however, the selection and training of the interviewers is critically important to accurate measurement.
On-site metering, use of other types of measuring equipment	High	High	Energy-use metering is sometimes used in outcome and impact evaluations and often used in evaluating the energy savings for specific buildings. Industry-developed guidance for the use of metering for such purposes is published in the International Performance Measurement and Verification Protocol(s) (www.ipmvp.org) and for Federal buildings in FEMP's M&V Guidelines (http://energy.gov/eere/downloads/mv-guidelines-measurement-and-verification-federal-energy-projects)
Building simulation modeling	Medium-to-high	High	Whole-building simulation is more often used to assess alternative building configurations relative to an energy use goal or to diagnose compliance with efficiency standards. Occasionally such information may be input to program evaluations.
Utility billing data	High, provided the evaluator understands how to get the desired information out of utility billing files	Low	Energy-use histories for specific energy customers of energy utilities may have relevance for evaluations (e.g., an evaluation of the Weatherization Assistance Program). Typically the request is made to the utility by a third party such as a state energy office. The utility must agree to provide the data voluntarily, and privacy issues may be involved. It can be challenging to understand and process utility billing files.
Mail surveys	Medium	Medium, usually higher than telephone	Non-response is an issue, although methods exist for compensating for non-response (see multi-mode methods below). The accuracy of responses about objects (e.g., recognition of logos) with whose names the respondent may not be familiar can be improved by providing a picture or drawing.

Telephone surveys	Medium	Medium	Non-response is an issue, although methods exist for compensating (see multi-mode methods below). Telephone interviews usually take less time to complete than the other methods because there is more control over the rate of response. The validity of responses to complex questions is a serious issue. If call lists must be purchased from a list vendor, the independent third-party evaluator will need to provide the vendor with documentation that the lists will be used for research purposes to avoid conflict with the National Do Not Call Registry. https://www.donotcall.gov/FAQ/FAQDefault.aspx
Website or email surveys	Medium to Low	Low	Non-response is an issue, although methods exist for compensating. The principal source of weakness is obtaining a probability sample so that statistical precision can be claimed. If many members of the population of interest do not have access to computers, it is difficult to claim a probability sample. Many Web or e-mail surveyors demonstrate that their respondents represent the overall population and make the claim. This method is growing in popularity; however, care must be taken to demonstrate that the population that is capable of being sampled is, in fact, the population of interest.
Interviews	Low	Medium	As used here, "interview" means the collection of data through protocols that will not support statistical inference. These are informal one-on-one question-and-answer sessions, usually with small numbers of respondents, which are designed to gather insights from experts on particular topics. Interviews can be conducted in-person, by telephone, or e-mail. See also "Focus groups."
Focus groups	Can make a defensible contribution to process evaluations, but otherwise low	Low	Focus groups are used to probe selected respondents in-depth for their reasons for a choice or their opinions regarding an event, process or object of interest. The findings of focus groups do not have statistical precision because the samples are very small (8-12 persons) and are usually not done using statistical sampling methods. Thus findings cannot be generalized.
Observation (e.g., mystery shopping)	Can make a defensible contribution to process evaluations, but otherwise low by itself	Low The cost of mystery shopping can increase to medium if travel for national-level research is required.	Mystery shopping is used to estimate participants' adherence to program rules without the participants' awareness of being evaluated. Usually the samples are non-probability samples. Observations are usually used as tools in process evaluations.

Literature review	Depends on purpose to which put, low to medium	Low	Literature reviews may contribute to *meta-evaluations* (borrowing the results from evaluations of comparable programs that operate under similar circumstances, synthesizing the findings, and applying them to the program) or obtaining anecdotal information (e.g., to use as evidence of external influences). Literature review may also be used to expand one's knowledge about the latest program theory for the purpose of developing effective evaluation questions.
Program records and reporting	Depends on purpose to which put, low to high	Often lowest	Program records and reports often serve as sources of data for process and output metrics. As such, they may be the most accurate data available, which can contribute to the evaluation's strength. If the accuracy of their data is questionable, however, it can weaken the evaluation.
Multi-mode methods (use of more than one method for a single data-collection)	Usually higher than the individual methods used alone	Higher than the individual methods used alone; however, synergy may help to reduce cost.	Combinations of mail, telephone, and in-person data collection for a particular survey can increase response rate and help to evaluate bias due to non-response. OMB sometimes expects the use of multi-mode methods for these purposes.

Secondary Data

The following are examples of secondary data sources:

- The U.S. Energy Information Administration (EIA) energy end-use data
- Census data
- Energy savings coefficients (i.e., estimates of energy savings [e.g., kilowatt hours] per unit outcome [e.g., installation of an efficient motor]) that were developed for one DOE program and may have relevance to another.[34]

If applicable secondary data are available, it is advisable to use them to supplement routinely collected data and other primary data because secondary data will significantly reduce data-collection costs. However, two very important caveats must be considered. The secondary data must be relevant and their transfer to the program for which they will be used must be defensible (see Appendix H for "Lessons Learned for Improving the Quality of EERE Evaluation Studies"). In particular, if the evaluation plans to use energy-savings coefficients or gross or net estimates from the evaluation of another program, the DOE evaluation project manager must ensure that the circumstances under which they were developed and the method of developing them are appropriate to the purposes of the current program's evaluation. Among other considerations, an energy-savings estimate would need to fit end-user industry and size profiles, as well as the

[34] At the time this Guide was being written, DOE was creating the Uniform Methods Project for Determining Energy Efficiency Program Savings to develop common energy savings coefficients for wide use by states, utilities, and the federal government. If the independent third-party evaluator must use standardized energy savings coefficients, this is an additional source to consult. http://energy.gov/eere/about-us/initiatives-and-projects/uniform-methods-project-determining-energy-efficiency-program-savings.

application profile, to be credibly applied to other end-user populations and technology applications.

Data Quality Control ("Data Cleaning")

The data should be examined for missing measurements, inconsistent measurements, and unusual values (outliers) prior to analysis. The evaluator's data collection plan should describe the quality control protocols it will use to examine the data for anomalies and the steps it will use to correct any apparent errors found. Data cleaning can be time consuming but it is very important for accurate information.

4.4.2 Census or Sample?

Another data-collection choice involves whether the evaluation collects data from: (1) the entire population of participants (like a census); or (2) from a sample of the population. Either option may be used for any type of evaluation; however, like most of the other choices, the choice has implications for cost and defensibility of the results. Table 4-3 highlights these options.

It will be very useful when communicating with evaluation experts to be aware of the difference between "***statistical precision***" and "***accuracy***" as used in survey-based data-collection activities. "Statistical precision," also known as "sampling error," applies to samples and consists of two parts: (1) how close (within a plus or minus interval) one wants a sample estimate to be to the true population value, and (2) the probability of getting a sample whose results will lie inside the desired interval. The former is the width of the interval within which the true value of the variable being estimated lies in relation to the estimated value (e.g., plus or minus 10%). The probability of getting a result that will lie inside this interval is the "confidence level" that the sample will deliver a result within this interval. Usually, "statistical precision" and "confidence level" together are specified as a "confidence interval" (e.g., +/-10% with 90% confidence, or often, 90 +/-10%). If statistical results are desired for any of the specific questions, a DOE evaluation project manager may ask the independent third-party evaluator to recommend the target confidence interval(s) for the findings.

Table 4-3. Options for Selecting the Number of Respondents from Which to Collect Data

Option	How Many Are Measured and Resulting Statistical Precision of Estimates	Rank Order of Contribution to Defensibility*	Relative Cost
Census	**Measure entire population.** Statistical precision is not applicable because every outcome is counted and, therefore, there is a full rather than partial enumeration. However, if the evaluation will measure the entire treatment and control groups for a RCT, this will be the option used.	Highest	Usually Highest
Sample Probability sample: Simple random and stratified random Systematic	**Measure a randomly selected subset of the population.** Probability of a population unit entering the sample is known. Sampling precision depends on the number of units (e.g., participants, measured). The more measured, the better the precision. These options are desired for quasi-experimental research designs.		Medium The cost will increase with the sample size.
Any non-random method of sampling	**Measure a non-randomly selected subset of the population** Probability of selection is unknown; statistical precision is not applicable. Carefully selected representative samples are sometimes claimed to have properties "similar to" probability samples.	Lowest	Usually lowest

"Accuracy" refers to the correspondence between the measurement made on an indicator and the true value of the indicator. Accuracy describes the exactness of the measurements made in the sample. In the sampling literature, accuracy is part of the concept of "non-sampling error." Accuracy should be a concern when the data-measurement instruments are designed. The independent third-party evaluator should always pretest questionnaires and other data-collection instruments before deploying them for actual evaluation measurements.

A common issue associated with taking a census and sampling is non-response (i.e., the fact that you will not be able to obtain information from some members of the population selected for your survey [unit non-response] or that those who respond do not answer all of your questions [item non-response]). Non-response threatens the strength of the results. The usual method of easing this threat is to require the evaluation contractor to demonstrate that those in the census or sample who did not respond to a survey are similar to those who did.

4.4.3 OMB Clearance to Collect Data

If the audience from which you need to collect data does not consist exclusively of Federal Government employees, and the evaluation needs primary data from ten or more members of this audience, including potential audience, then the data collection activity will require approval (clearance) by the OMB under the legislative requirements of the Paperwork Reduction Act (PRA).

Paperwork Reduction Act's (PRA) Key Relevant Points

The Paperwork Reduction Act (PRA) prohibits federal agencies from conducting or sponsoring a "collection of information" without prior OMB approval (5 CFR 1320.5). Collection of information refers to any instance of consented data gathering from ***ten*** or more persons, regardless of whether the data gathering is "mandatory, voluntary, or required to obtain or retain a benefit" (5 CFR 1320.3(c)). Although the law specifies that "government-owned contractor-operated facilities, including laboratories engaged in national defense research and production activities" are not "agencies," (5 CFR 1320.3(a)(4), it notes that a collection of information conducted by such organizations may still trigger the PRA if a federal agency "sponsors" the collection. What constitutes sponsorship includes: "(1) causing another agency to collect the information; (2) contracting or entering into a cooperative agreement with a person to collect the information; (3) requiring a person to provide information to another person; or (4) in similar ways causing another agency, contractor, partner in a cooperative agreement, or person to obtain, solicit, or require the disclosure to third parties or the public of information by or for an agency (5 CFR 1320.3(d). When an agency "sponsors" a collection of information, OMB approval is required.

Federal government employees are excluded from the OMB clearance requirement only if the questions to be asked of them involve activities associated with their employment; otherwise, surveys of federal employees (e.g., as civilian participants in a program) also require OMB clearance.

The time required to obtain OMB clearance varies:

- For customer satisfaction surveys and pretests of other survey instruments, DOE has an expedited process that, in most cases, takes two to four weeks.[35] The Forms Clearance staff of EIA's Statistics and Methods Group can assist EERE staff with this process.
- For surveys other than customer satisfaction surveys, the OMB clearance process takes longer. Currently, the entire clearance process may require five to eight months. EERE clearance applications are submitted to the Records Management Office (IM-11) of DOE's Chief Information Officer.

An OMB clearance is valid for three years.[36] Appendix D contains additional information about how to obtain an OMB clearance for data-collection.

[35] OMB approved generic clearance DOE-887.
[36] Clearances resulting from emergency reviews last six months; however, emergency reviews are unlikely to apply for general evaluations.

4.5 Choose Appropriate Analytical Method(s) for Selected Research Design

Analytical methods for a selected research design are needed to: 1) analyze and develop findings from the research data; and 2) treat threats to validity, as necessary.

Many analytic methods are available for developing findings from data. Table 4-4 provides a brief overview of some of the more common analytic methods used to develop evaluation results. If the data were collected by a probability sample, select analysis methods that can make use of the statistical properties of the sample. These methods are identified in the "Typical Use in Evaluation" column of Table 4-4.

Many of the methods described in Table 4-4 can be used for more than one type of evaluation. The types of evaluation in which the methods are more commonly used are indicated in the "Comment" column. More than one of the methods listed may be used in the same evaluation analysis. For example, use of deemed energy savings or engineering analysis is sometimes used to create an estimate of the energy saved from installing a particular energy conservation measure.

Table 4-4. Common Analysis Methods Used for Program Evaluations that are the Focus of this Guide

Analytical Method	Typical Use in Evaluation	Defensi-bility	Relative Cost	Comment
Case study	Describe the causal chain leading to an outcome. They are typically used in process and impact evaluations, including R&D program evaluations and deployment success stories.	Low to medium	Low	This is an option if the budget is tightly constrained; however, the ability to deduce defensible findings is usually weak. The latter, in particular, may be valuable for attracting additional participants.
Content analysis	Identify themes that exist in unstructured data (e.g., identify the most frequently sought information from inquiries to a call center, or find the themes in focus group transcripts). They are typically used in process evaluations.	Medium	Low to high	The cost of a content analysis will depend on the number of concepts found that are relevant to the evaluation objectives and the number of data sources that have to be content-analyzed. If the number of sources is large, computer algorithms exist that will help to manage costs.
Meta evaluation: evaluation synthesis	Synthesize the findings from evaluations of similar programs that operated under similar circumstances and use them as findings for the program being evaluated. The synthesized findings may also be used as a benchmark for the program being evaluated. They are typically used in process and impact evaluations.	Low	Medium	Meta evaluations can be labor intensive. It may be costly to search for, assess the relevance of, and extract the relevant findings of other evaluations. The programs whose evaluations are reviewed must be similar to the program under evaluation and their evaluation findings must be relevant to the current evaluation's objectives.
Expert judgment Delphi analysis	These forms of expert judgment can be applied in circumstances where (1) collecting quantitative data might be very difficult or costly, and (2) experts exist	Low to high (wide range)	Low	Delphi analysis is a systematic collection, comparison, and synthesis of judgments from several experts on a subject (e.g., the amount of an outcome that is attributable to the program). If the experts cannot reach agreement on a finding, however, the

Peer review	who are willing to support the evaluation. They are typically used in process and impact evaluations.			process may be severely discredited.[37]
Cost-benefit analysis	Link program achievements to resources expended.	Low to high	Low to high	Usually, cost-benefit analyses are quantitative. At a high level of program aggregation the evaluation cost is low and its strength is good because quantitative cost data are usually available and direct benefits can be estimated with less effort. But if the analysis is for disaggregated activities that are part of an overall program strategy, it may be so difficult to disaggregate the costs and benefits to the activity level that the results are open to challenge. The benefits of indirect effects (e.g., gaseous emissions reductions and national security), may be difficult to quantify credibly.
Engineering estimation	Calculate estimates of energy savings or emissions reductions based on engineering, physical, and chemical theory. They are typically used in impact evaluations.	Medium	Low to medium	Usually calculated as an average for a set of circumstances encompassing those encountered by the program, then stipulated for all similar circumstances. Energy savings coefficients are often developed by engineering estimation.
Tabulation & cross-tabulation	Count activities, etc., and place them in categories of interest. These are typically used in process and impact evaluations.	Medium-to-high Depends on use	Low	Tabulations are used to report the number of outputs, outcomes, etc., observed. Cross-tabulations report the number of outputs, etc., that occur jointly in two or more categories of interest.
Correlation	Statistically estimate the strength of a relationship between two indicators. They are typically used in both process and impact evaluations.	High	Low	Used to determine the degree of relationship (covariance) between selected output and outcome indicators or any two variables.
Regression, including econometric and discrete choice analysis	Statistically estimate an equation that calculates the value of an *outcome indicator* (e.g., energy saved), given the value(s) of one or more *output*, *activity*, or *external-factor indicator(s)* used as an independent variable(s) (e.g., receipt of a training class, installation of an energy-efficiency measure, energy price). They are typically used in impact evaluations.	High	High	A regression equation that includes variables for the known influences on energy usage can estimate the amount of energy saved by program participation (one of the influences). The significance of the regression coefficient for the variable representing participation indicates whether the resulting estimated value of the outcome (e.g., savings per unit) of participation is statistically significant and, therefore, defensible. If data for non-participants are included in the regression analysis, the coefficient of the participation

[37] Delphi analysis (a method of decision-making and forecasting that involves successively collating the judgments of experts) must be used in the context of the Federal Advisory Committee Act's (5 USC App 2) restrictions on directing Government-formed advisory committees to reach a consensus, as described in EERE's "Peer Review Guide," July 2004, p. 23. These restrictions do not apply, however, if an independent third-party evaluator establishes the panel. In many cases, the members of a Delphi panel work independently and are not necessarily directed to produce agreement on a finding.

				variable may be interpreted as net energy savings. Many variants exist for this method. An evaluation expert proposing to use this method for an evaluation should provide evidence of expertise in its use. For maximum defensibility, an independent QA review of the expert's evaluation plan is advised.
Differences of means and proportions	Comparison of two or more groups on an indicator of interest	High	Low	Can be used to compare two groups (e.g., a participant and non-participant group), on how their behavior changed on an indicator of interest during the program period. The analysis should include a test of statistical significance. Typically used in process and impact evaluations.
Survival Analysis	A statistical modeling method used to evaluate persistence of energy savings	Medium	High	Several models of survival analysis exist. A qualified statistician or econometrician is usually required to choose among them. Typically used in impact evaluations.

4.6 Participate in an External Review of the Evaluation Plan

At this point, the evaluator has prepared the evaluation plan for review. The DOE project manager must manage the external and internal peer review process and create a peer review panel. It is of vital importance that the review panel is chosen with great care, ensuring that it is filled with experts with the right balance of skills for the study to be done.

The quality assurance/peer review protocols described in Section 3.3, Step 2, must be implemented when the evaluator's draft evaluation plan is submitted to the DOE evaluation project manager. The evaluator is expected to:

- Read and agree to the Review Charter (see Appendix G for a sample Review Charter).
- Send the evaluation plan to the DOE evaluation project manager, allowing a reasonable period of time for the reviewers to provide written comments.
- Participate in a review meeting with the external reviewers. Discuss and clarify the reviewers' major critiques and proposed resolutions. If the program under evaluation is a high profile or high priority program, or has a large budget, the meeting should be in-person. For lower priority, or low-budget evaluation studies, a conference call will suffice.
- Modify the evaluation plan in response to the critiques.
- Once the DOE evaluation project manager and reviewers (as necessary) have had a chance to see and approve the proposed modifications, the evaluation project manager can approve the plan for implementation.

5.0 STEP 4. CONDUCT THE EVALUATION

Once the evaluation plan has undergone a formal peer review and is modified to the satisfaction of the review team and evaluation project manager, the evaluator can proceed to implementation. Here is where the systematic effort to collect the primary data for the evaluation is done. The data are quality-assured on a continuous basis during collection, and then cleaned for analysis. The analyses and calculations are then done to answer the evaluation questions.

5.1 Perform Sampling, Data Collection, Measurement and Verification

5.1.1 Sampling

Sampling allows data about a population to be collected at a reasonable cost by gathering them from a subset instead of the entire population. Almost without exception, an evaluation study would require the use of sampling because it would be cost-prohibitive to conduct a census. Important steps include defining the measurement objectives of the study, determining the sample size requirements, and selecting the sample.

The measurement objectives usually cover questions on what is to be measured, from whom, and at what level of precision. Factors influencing sample size decision include the number of measurement units in the target population, the baseline level of the indicator, the level of statistical significance and the degree of statistical power. Finally, selecting the sample usually involves the use of probability (random) or non-probability (non-random) sampling, depending on the type of study that is possible. The default requirement is for evaluators to use probability sampling as long as that is an option, and to only use non-probability sampling methods if probability sampling is not possible for the study.

Some examples of probability sampling are simple random sampling; stratified sampling; cluster sampling; systematic-random sampling; and multi-stage sampling. The evaluator would decide on which approach is appropriate for the purposes of the study to be done, and justify the choice in the evaluation plan for comment from the review panel, ensuring that there is agreement about what constitutes the best sampling approach to meet the needs of the evaluation to be done.

5.1.2 Data Collection

Two essential elements of data collection, instrumentation and measurement approach, require attention to ensure that the collected data are defensible.

- **Instrumentation**. An instrument can be a survey tool, in-depth interview guide, or a metering device or other technological measuring device. Often it is a combination of these. Measurement validity depends both on instrument validity (is the instrument measuring what it is supposed to be measuring?) and reliability (are the measures consistent over repeated takes) of the measurement approach.
- **Measurement approach**. The measurement approach (i.e., *how* the measurement is done) must be transparent and reliable to maximize the gathering of defensible information and support the validity of the findings.

5.2 Complete Data Analyses and Calculations

Analysis is the next step once the evaluation data have been gathered and properly entered into a database that can be read by the data processing tool used for the study (usually a spreadsheet or commercial statistical analysis software). Prior to data analysis, all data must be properly cleaned and quality checked using the protocols specified in the evaluation's data collection plan.

Above all, the actual analysis must be guided by the evaluation question(s). The analysis performed should be faithful to the analytical strategy approved in the evaluation plan. The analytical procedures must be transparently documented, and the underlying database made available to EERE, as per contractual agreement.

Although the specific analysis would depend on the aims of the evaluation, if the evaluation inquiry involves some of the common metrics often assessed in EERE evaluations, it is imperative that the evaluator refers to existing EERE guides on acceptable minimum standards for the treatment of those metrics.

5.3 Identify Key Findings

The answer to each evaluation question must be clearly identified and demonstrated for transparent reporting. There should never be an attempt to deceive, either through commission or omission, the findings of the study. It is not uncommon for evaluators, despite their independence, to feel some unspoken obligation to exaggerate the strength of positive findings or to minimize negative findings. Care must be taken to avoid such. If there are weaknesses in the evaluation, they should be caveated, but the caveats shouldn't be seen as overwhelming any value provided by the evaluation.

It is critical to identify key findings for each evaluation question so that, prior to forming the narrative around the results, it will be possible to answer the question of impact and process for high-level stakeholders who often need a bottom line response to their inquiries. Additionally, such transparent identification of key findings provides the fulcrum around which the overall findings are reported.

6.0 STEP 5. MANAGE IMPLEMENTATION OF EVALUATION PROJECT

Managing the evaluation project during implementation is as much art as science. Some of the concrete activities include holding and participating in periodic project progress-review meetings, reviewing and acting on project status reports, monitoring independent third-party evaluator achievement of milestones against expenditures, and managing the internal and external review process. Through all of these efforts, the manager must maintain the capacity to resolve ongoing and evolving issues that the evaluator requires assistance from DOE to navigate. The DOE evaluation project manager must have the management skills to keep the project from veering off course.

6.1 Hold and Participate in Periodic Project Progress-Review Meetings

The actual implementation of an evaluation study is typically fraught with unanticipated obstacles. This can lead, even in the best of circumstances, to delays in schedule; in the worst of circumstances, it may derail the study altogether. A key strategy for avoiding the pitfalls of implementation is through progress review meetings.

These meetings should be contractually required for every evaluation study and cover the entire implementation phase. A regular schedule must be established and kept, although the medium of the meetings (e.g., telephone conference or video conference) should reflect what is most convenient and least burdensome for all parties. Any deviations from the schedule would need the mutual consent of the DOE project manager and the evaluator, and occur only in the interest of benefiting the successful conduct of the study.

> The most carefully designed evaluation project can fail to provide defensible information if it is not monitored.

Progress review meetings are aimed at ensuring that project activities progress according to plan, and that obstacles that arise are identified and addressed in a timely and efficient manner, to ensure that a successful evaluation is conducted. They are indispensable to the process of management.

6.2 Review Project Status Reports from the Independent, Third-party Evaluator

Monthly project status reports, written and submitted by the evaluator, are useful monitoring tools for both the evaluator and for the evaluation study manager. The status reports need not be overly detailed or burdensome. A boilerplate template of no more than 1 – 1½ pages might suffice. The information they cover should include:

- What was expected to be accomplished in the preceding month
- What was accomplished during the preceding month
- What is expected to be accomplished in the next month
- Any special issues requiring resolution.

The DOE evaluation project manager has the responsibility to review and act on any issues requiring resolution. The monthly project status reports combine with the progress review meetings to provide two of the key monitoring tools for ensuring the success of the evaluation. They can be labor intensive, but they are critical.

6.3 Monitor Independent, Third-party Evaluator Achievement of Milestones and Expenditures

The third critical monitoring tool in the arsenal of the DOE evaluation project manager is matching milestone achievements with completion target dates and expenditures. Performing this function requires use of the budgetary breakdown for the project submitted as part of the contract, the timeline for conducting the study, and the monthly project status reports.

The timeline provides an overview of the distribution of the project expenditure over time, and thus offers a rough guide for what the project balance should be relative to outstanding tasks. For example, it is generally the case that data collection is one of the most labor-intensive aspects of conducting a study. Thus the resources that the evaluator commits to the background research leading to the development and review of the evaluation plan should be monitored to track how much of the project fund is being consumed, given that the bulk of the effort generally comes after the planning phase of the evaluation process.

The following are some of the specific aspects of this monitoring effort:

- Monitor the timeliness of the independent third-party evaluator's achievement of milestones.
- Monitor independent third-party evaluator invoices relative to work progress. The rate of expenditure on the project can provide an early warning sign of problems to come.
- Review all milestone products.
- Meet all of the contractual milestones for deliverables or support that have been promised to the independent third-party evaluator.
- If any of the foregoing elements deviate from expectation, conduct a special progress review meeting to confront the issue and develop a resolution before it becomes a serious problem.

6.4 Manage the Internal and External Review Process

The review process is critical for assuring the quality of an evaluation. For all EERE program impact and process evaluations, peer reviews occur at two milestones: A) following the development of the evaluation plan; and B) following submission of the draft evaluation report. In EERE, the peer review procedure, in brief, covers the following steps:

- The evaluation manager engages the team of independent peer reviewers for the review.
- The independent third-party evaluation team submits the document to be reviewed (evaluation plan or draft report) to the project manager, to be passed on to the reviewers.
- The independent panel reviewers submit their comments independently to the DOE evaluation project manager, who serves as the review manager. They may also submit redline edits to the document.
- The chairperson (if there is a designated chairperson for the review team) distills the key critiques, and includes it with all of the individual review comments for submission to the

DOE evaluation project manager and the evaluator; if there isn't a chairperson for the review team, this function is fulfilled by the DOE evaluation project manager.

- The evaluator and peer review team hold a clarifying meeting; this may be in the form of a conference call.
- The evaluator prepares a response to each critique, which is submitted to the peer review team and the DOE evaluation project manager.
- The evaluator proceeds to modify the product (evaluation plan or draft report) accordingly. If the peer review team and/or DOE evaluation project manager is satisfied that their concerns will be remedied, approval is given for implementation to begin.

These steps are repeated for an internal team of reviewers as well, with the exception that the internal team cannot be deemed to be without conflict of interest. While the internal review team is generally expected to provide useful, programmatic context to their review, they cannot be allowed to force or request changes to the findings of the study that in any way lessen the integrity of the analysis.

6.5 Anticipate and Address Technical and Management Challenges

As noted above, no matter how well a study is planned, there are almost always unanticipated obstacles that affect the plan. Sometimes the obstacle may even be anticipated, but still prove difficult to deal with.

The best way to overcome this sort of crisis is to anticipate as many of the pitfalls as possible and prepare contingency plans. For example, a study might propose a particular method to answer an evaluation question, contingent on acquiring the necessary data. If the needed data were unattainable for any reason, it would be impossible to complete the study. All stakeholders, including the evaluation management team, the independent third-party evaluation team, and the members of the review panel, must consider the possibility that the necessary data might not be available, or obtainable at the desired level of thoroughness and quality, and propose back-up plans in case of such eventuality.

Evaluability assessments can serve as a key tool to guard against unanticipated obstacles, but even these cannot always presage what will actually transpire during study implementation. By combining rigorous use of the monitoring tools with the quality assurance embedded in the peer review process, the project manager enhances the likelihood of a successful evaluation study with valid, defensible results.

7.0 STEP 6. REPORT THE EVALUATION RESULTS

The evaluator is expected to prepare a draft and final report at project completion using DOE reporting principles and guidelines[38], participate in a peer review of the draft report, and then prepare and publish the final report. The DOE evaluation project manager ensures that these tasks are executed to meet the standards of scientific reporting.

7.1 Prepare Draft and Final Evaluation Report

An evaluation study results in a report written by the evaluator. The major content of the report consists of the answers obtained for the evaluation questions.

It is important to specify before the evaluation begins the types of information that must be in the evaluation report so that its content will serve its intended uses.

Appendix D (1&2) contains examples of EERE evaluation report outlines. In general, a report outline includes sections for the following:

- An Executive Summary
- Background of the program or intervention to be evaluated, its logic and the objectives for the evaluation
- A description of the research design and assumptions, how the data were collected, the analysis method(s), and the limitations of the study. These descriptions should be brief in the main report, where the focus is on the answers to the evaluation questions and recommendations. Put the comprehensive, technical expositions in an appendix.
- Answers to all of the questions specified for the evaluation
- Recommended improvements to the program, if relevant (indicate which are high priority compared to others)
- Recommended improvements to the evaluation process that address limitations of the analysis, as well as any lessons learned about data collection and analysis methods that might aid future evaluations. These can be based on the evaluator's experience and observations during the evaluation process.

7.2 Participate in Peer Review of Draft and Final Evaluation Report

Section 6 provided an overview of the peer review process for the evaluation products. This process is deployed for the review of the draft evaluation report. It is the responsibility of all involved stakeholders – the evaluator, the DOE evaluation project manager, and the peer review team – to work to ensure that the required evaluation report is of the quality and high standards of DOE and the professional evaluation community.

[38] EERE reports follow the general format of a journal article. However, an Executive Summary takes the place of an Abstract. The remaining sections are more conventional: Introduction / Background; Methods; Results; and Conclusions.

8.0 STEP 7. USE THE EVALUATION FINDINGS

After the evaluator submits the final report, there are steps that help to ensure the evaluation is used and its purpose satisfied. Additionally, the data that formed the basis of the analysis for the impact and process evaluations should be stored in EERE's repository database of impact evaluation results.

8.1 Distribute the Evaluation Report and Results

The DOE evaluation project manager has the responsibility of ensuring that the evaluation report gets into the hands of those who need to use the results. This consists of sending copies of the evaluation report(s) or notices of their availability to all of those who can use the results.

Ensure that those who need to use, or can use, the evaluation findings receive them.

This audience is not limited to the decision-makers who are waiting for the results to inform their decisions. It should also include the public, as well as stakeholders in the program and other program staff who might benefit from the findings.

Finally, it is recommended that an action plan for disseminating the evaluation findings should be developed during the final review of the draft study report. An action plan might include:

- Making presentations to decision makers, program staff, implementers, and stakeholders; and
- Sharing lessons learned about the evaluation process with other DOE evaluation project managers.

In sum, it is good management practice to be proactive about getting the results noticed and utilized. The publicity given to the findings may help the program.

8.2 Use the Results to Make Decisions about the Program

The objectives that motivated the conduct of an evaluation will indicate the uses for the results. Broadly speaking, an evaluation's uses consist of (1) demonstrating ***accountability*** (disseminating the results through purposeful, targeted communication to show the program's value) and (2) ***learning/improvement*** (using the results internally to make decisions about improving or otherwise changing the program, taking care to communicate and motivate). These uses will also help the program manager decide who must receive the evaluation's results.

Once the evaluation study is finished, use of the results must begin by answering the question: "Why did we decide to do this study in the first place?"

Some general questions to ask once the evaluation results are in hand:

- Did the findings provide evidence that previous investments have achieved the intended positive benefits?
- Did the findings suggest that the program has been achieving unintended effects? Are the unintended effects positive or negative?
- Did the findings provide evidence as to which specific aspects of EERE's investments were most responsible for the impacts (if this was part of the research questions)?

- What was the scope and scale of the impacts? Did it justify the investment?
- What lessons do the findings provide with respect to ongoing investments? Do the findings point to a need to explore unexamined aspects of current investment? Do they suggest the need for even more information?

Although impact evaluation studies are not aimed at in-progress modifications (process evaluations serve that task better), findings from impact evaluation can serve several purposes related to program improvement. These include, in addition to understanding whether intended impacts were achieved, what unintended impacts might have also resulted, if these were negative, and whether any new investments might be justified. For that matter, the findings can provide for a better understanding of the specific aspects of the previous investment strategies that facilitated program goal achievement.

8.3 High Impact Communications

Effective communication of the impacts achieved by the program represent another dimension in the use of evaluative information. The DOE evaluation project manager must avoid the mistaken impression that the results would sell themselves, even when the results are spectacular. The core guiding principle in communicating the results of impact evaluation results is to be accurate. However, successful communication – that is, whether the communication achieves what was intended – involves more than simply being accurate. It requires careful planning, and the use of principles of communication (to develop the messages) and marketing (to distribute the messages).

Marketing principles, in particular, provide useful guidance for EERE to successfully communicate the results from impact evaluation studies. Successful marketing of the program's impacts, as described by the evaluation results, entails the following:[39]

- **Market segmentation**. Divide the stakeholders into distinct audiences who have different needs, characteristics, or behaviors, and who might require separate messages.
- **Market targeting**. Evaluate each stakeholder segment's importance as if they are a market and determine which ones deserve to be prioritized for communication of a given evaluation study.
- **Positioning**. Position the message about the results to occupy a clear, distinctive, and desirable place, relative to competing messages.
- **Differentiation**. Differentiate the message relative to competing messages to create a perception of superior value in the minds of target stakeholders.

The communication strategy in this guidance has the purpose of facilitating stakeholder value and providing the appropriate platform for profitable customer/stakeholder relationships. Whatever the strategy chosen, it is imperative that the findings are properly understood through an accurate communication of the results, without exaggeration or misrepresentation.

[39] Kotler P and Armstrong G. (2012). "Principles of Marketing." New Jersey: Pearson Prentice Hall.

8.4 Establish/Update Program Records For Use in Future Evaluations

Each program evaluation is an investment of program resources to formally assess program achievements and identify recommendations for program improvements. The evaluation experience itself is valuable because it shortens the learning curve for the next evaluation. To preserve this knowledge and associated learning for future use, the DOE evaluation project manager should:

- **Ensure that evaluation data relevant to the program are being gathered centrally, and in a systematic way.** Routinely collected data should be warehoused in a central data system. Funding Opportunity Announcements (FOAs) and subsequent awards should require the routine reporting of quality-assured evaluation data for use in future evaluations.
- **Establish a database** for warehousing evaluation results data and recommendations.
- **Archive discussions of the problems encountered** in the evaluation and how they were resolved.
- **Archive the updated contact information** on the individuals and organizations contacted during the planning and implementation of the evaluation, while protecting any required confidentiality.

APPENDICES

Appendix A. Example of Statement of Work for an R&D Evaluation Study

Statement of Work (SOW)
Subcontract Number ____

Background and Evaluation Focus
DOE's EERE benefit-cost studies seek to capture benefits and costs across four dimensions of import to the agency and to society, broadly: economic, energy, environmental and energy security. The key evaluation questions to be answered by the retrospective benefit-cost evaluation are:

- *To what extent has DOE/EERE produced net economic benefits (resource savings and other positive economic effects) relative to the next best alternative?*
- *To what extent has EERE achieved energy and environmental benefits and enhanced energy security?*
- *To what extent has EERE cultivated a knowledge base in the research community that has impacted innovations in today's markets?*[40]
- *Would today's commercialized technologies likely have happened in the same timeframe and with the same size of effect without EERE 's efforts?*
- *To what extent do benefits attributable to EERE involvement exceed EERE expenditures? Was the public investment worth it?*

The benefit-cost analysis identifies a DOE cluster of evaluative interest (i.e., a DOE research area or sub-program or program), and then selects a few technologies/projects within the cluster for detailed analyses while the rest are treated qualitatively. Combined benefits of the technologies evaluated in detail are compared against the cost of the entire program or sub-program or research area (depending on how the cluster is defined). Economic performance metrics that are calculated are net present value benefits, benefit-to-cost ratio, and internal rate of return. Note that these performance metrics are to provide estimates of the return on DOE's investment in (1) the selected technologies, and (2) the larger defined cluster. This return may be all or part of the larger societal return from having, versus not having, the technologies, depending on the attribution to DOE. The economic performance metrics are calculated using two elements of monetized benefit as described below: "economic benefits" and "monetized health benefits." Sensitivity analysis includes examining net present value and benefit-to-cost ratios with both 3 and 7 percent discount rates, as well as sensitivity analysis for other uncertainty factors deemed critical to study findings.

Economic benefits are quantified by comparing actual technological progress to counterfactual scenarios. Energy, labor, and other resource requirements with the use of the selected technologies are compared against those had the next best alternative technologies been used instead. In addition, the analysis identifies the portion of the difference between actual and counterfactual that can be credibly attributed to the DOE. While these two parts of the analysis - comparison with the next-best alternative and assessment of additionality - are in practice often combined, explicit treatment of each is requested here.

[40] TIA Consulting will conduct this activity.

For example, a new material may offer cost savings compared with the next best alternative material that would have been used had the new material not been developed. Assessment of additionality may determine, for instance, that 80% of the advancement was attributable to DOE's involvement and 20% of the advancement would have happened without DOE. Thus, 80% of the estimated cost savings are taken as cluster economic benefits. Only if 100% of the difference between the selected technologies and the next best alternatives are attributed to DOE's funding is the total difference taken as an economic benefit of the cluster.

Environmental benefits quantified will focus on reduction in air pollution. Monetized health benefits can then be calculated using EPA's Co-Benefits Risk Assessment (COBRA) screening model. Energy security benefits will focus on cumulative fuel displacement (e.g., oil and natural gas), expressed in physical terms. Important benefits that cannot be measured quantitatively within the scope of a study are discussed qualitatively.

Knowledge benefits will be identified by TIA Consulting under a concurrent contract, in consultation with the LBNL Project Manager and the subcontractor leading this evaluation. The TIA Consulting study will use interviews, database review, and patent and publication citation analysis to shed light on the paths through which outputs of DOE-funded R&D have been disseminated to producers and users. This work will be an update to earlier work (Ruegg and Thomas 2008) and will be examining NiMH battery work separately from other battery R&D. This work will be part of the knowledge chapter in this report and will also be incorporated in a separate section of the report to support the discussion of attribution. TIA Consulting will be listed as a contributing author to the report.

TASKS

Task 1. Conduct a project initiation meeting with LBNL Project Manager, DOE Project Manager, and DOE staff to discuss proposed work and schedule.

The purpose of this meeting is to:

- Review and refine the research objectives and methods, discuss current data availability and next steps on determining data requirements and data availability and quality;
- Review and confirm the scope of work and deliverables with LBNL and DOE staff regarding the evaluation;
- Review and adjust (as necessary) the project approach outlined in the proposal; and
- Clarify program management and communication protocols.

In consultation with the LBNL Project Manager, the evaluator shall prepare an agenda for the meeting. The evaluator shall also prepare and submit to the LBNL Project Manager for approval a detailed memorandum documenting the results of the project initiation meeting.

Task 1 Deliverables:

- Agenda for project initiation meeting
- Memorandum documenting the results of the project initiation meeting

Task 2. Conduct preliminary review of key documents, hold meetings and interviews with program managers and key stakeholders, and conduct preliminary modeling and planning

In this task, the evaluators will conduct a critical review of key documents, including DOE' s Energy Efficiency and Renewable Energy (EERE) Methodology Guidelines. In addition, the

evaluator is expected to hold meetings and interviews with DOE program managers and other individuals familiar with NiMH battery R&D and investments. Working with program managers, the evaluator will identify the NiMH battery technologies to be studied in detail. The DOE will provide historical program and project cost data that will determine the level of cluster analysis that is feasible. The evaluator will also complete an initial investigation of these technologies, as well as preliminary modeling, and preliminary planning for a cluster approach.

Task 2 Deliverables:

- List of material reviewed
- List of people interviewed
- Initial plan for cluster approach, selected technologies, and next best alternatives Task 3.

Task 3. (A) Draft a detailed logic model laying out the underlying theory of change for the investments to be evaluated; (B) Incorporate the logic model in an evaluation plan for conducting the benefit-cost impact evaluation, and participate in a peer review process.

i. Based on the work conducted in the previous task, the evaluator shall review the program and market theories for the initiative, refine the theories as necessary, identify any alternative theories that merit review, develop the program logic model, specify program performance metrics, and identify data that the evaluators can expect to obtain from project grantees.
ii. Based on these models, the evaluator will prepare an evaluation plan for conducting the evaluation study, covering both process and impact evaluations.
iii. The initiative model and evaluation plan will be peer reviewed by a group of qualified and independent reviewers. LBNL will manage the peer review process, and the evaluator is expected to participate in the peer review process.
iv. After the peer review, the evaluator will revise the logic model and evaluation plan, as needed.

At a minimum, the evaluator shall include the following sections in the evaluation plan:

1. Research Objectives and Approach
    a. The research objectives as refined in Task 1
    b. A short description of the investments that are being evaluated
    c. A presentation of the program theory of change
    d. A short description of the expected program effects and process metrics
    e. A list of the services offered by the program with an indication of which services will be addressed in the evaluation
    f. A presentation of the researchable issues and questions to be addressed in the evaluation and reported in the evaluation report
    g. A presentation of the overall scope of the evaluation
    h. A detailed presentation of the evaluation activities to be undertaken
2. Data Requirements
    a. Identification and description of all data necessary to meet the research objectives and conduct the analysis
3. Secondary Research
    a. Identification and description of any secondary research that will be collected
4. Data Acquisition

a. Identification of people who will be surveyed / interviewed
b. A presentation of the sampling approach and sample selection methods for each evaluation activity that includes sampling efforts

5. Data Cleaning and Quality Control
6. Analysis Methodology
    a. A description of the data handling and data analysis approach to be used to address the researchable issues
    b. A description of how the comparison group or non-participant information will be used in the evaluation
    c. A presentation and discussion of the threats to internal and external validity, methods used to treat them, and the level of uncertainty associated with the sample selection methods and the evaluation approaches
7. Method for Treatment of Attribution
8. Schedule and Milestones
9. Final Report Outline
10. Project Management
    a. Detailed work plan and schedule for the study by tasks and subtasks
    b. Discussion of project management issues, including procedures for identifying and monitoring the progress of the evaluations
11. Budget
    a. A presentation of evaluation costs by tasks.

Task 3 Deliverables:

- Draft evaluation plan
- After peer review, revised evaluation plan

Task 4. Conduct data collection and analysis and provide interim feedback

The evaluator shall conduct the study using the EERE specified methodology that was developed, tested, and improved in four 2009-2010 studies and described in a draft guide and related publications (see References). As noted above, this study will include data collection, technology and market assessments, quantitative estimations of benefits and costs for selected technologies, assessment of next best alternative and attribution, computation of economic, energy, environment and energy security performance measures for the selected technologies, incorporation of knowledge measures provided to the study by a supporting effort, and any supporting qualitative treatment of selected technologies as needed.

No data request can ask the same questions of more than 10 people because there is not sufficient time for Office of Management and Budget (OMB) clearance.

The evaluator shall provide periodic feedback to the LBNL and DOE Project Manager about the status of the work and what appears to be working and not working. Feedback shall be in the form of findings memos and conference calls. The evaluator shall coordinate with TIA Consulting on the knowledge benefits analysis and reporting aspects of the overall study.

Task 4 Deliverables:

- Findings memos summarizing status of work and periodic feedback
- Conference calls with DOE, LBNL, to present and discuss
- Minutes from conference calls summarizing status of work and periodic feedback Task 4.

Task 5. Coordinate with TIA Consulting

The evaluator will coordinate with TIA Consulting for supporting analysis of knowledge benefits, and incorporate those findings into the benefit-cost study report. This work will be part of a knowledge chapter in this report, and she will be a contributing author to the report.

Task 5 Deliverable: N/A

Task 6. Prepare draft and final reports, participate in peer review process

The evaluator shall provide a draft report that, at a minimum, shall include the following sections:

- Executive Summary, emphasizing the major findings and the most significant recommendations.
- Background or Introduction chapter including the research objectives and description of the technologies.
- Methodology chapter describing and justifying the chosen approaches, data sources, data collection methods used in the evaluation, and any limitations of the data/analysis.
- Results chapter.
- Summary of results and Conclusions chapter.
- Appendices

The executive summary will be written in language accessible by lay persons both to the technology and the evaluation methodology. The body of the report will, however, be structured and written as a technical report. The draft report will be peer reviewed by a group of qualified and independent reviewers. LBNL will manage the peer review process, and the evaluator is expected to participate in the peer review process. After the peer review, the evaluator shall finalize the report (near publication ready) by incorporating any comments or changes (if applicable) from the peer review process. The evaluator will review the final publication report provided by DOE.

Task 6 Deliverables:

- Draft report
- Final report

Task 7. Brief DOE [Subject to Available Funds]

After the final report has been prepared, the evaluator will present the findings from the evaluation to DOE staff and other interested individuals.

Task 7 Deliverables: DOE briefing

Task 8. Provide Project Management

Under this task, the selected evaluator will be responsible for:

- Ensuring that all the evaluation work activities are implemented and that project

reporting is completed.

- Ensuring that the evaluator's contract management obligations are carried out in a professional manner.
- Managing sub-contractors, if any, so that the evaluation team speaks with one voice through the prime contractor.
- Maintaining regular and direct communication with the LBNL Project Manager.
- Maintaining and archiving electronic and paper files and data collected or developed during the conduct of the evaluation work. The documentation is the property of LBNL and will be turned over to LBNL at the end of the contract term. It must be in a form and structure that supports a chain of evidence for all evaluation findings.
- Attending and occasionally facilitating meetings - including initiation meeting(s), regular and ad-hoc project meetings, and a final evaluation "close out" meeting (via conference call).

Task 8 Deliverables:

- Electronic and paper files and data collected or developed during the conduct of the evaluation work

Task 9. Provide Project Status Reporting

Project reporting is expected to include, but will not necessarily be limited to, the following items:

i. Weekly or bi-weekly – On an agreed-upon day and time, the evaluator and the LBNL Project Manager shall have a weekly or bi-weekly conference telephone call, during which the evaluator shall provide a study update.
ii. Monthly – Project status reports highlighting issues with each evaluation activity and problems (e.g., including discussion of difficulties in getting the job done, coordination with other organizations, etc. with recommended or agreed solutions).
iii. Quarterly – Review of major findings, observations, project implementation and recommended updates to the evaluation plan.

Task 9 Deliverables:

- Monthly status reports
- Quarterly updates

Deliverables and Schedule:

The project is expected to be completed within 12 months after the start of the project. A detailed schedule for the deliverables will be developed as part of the first task. The schedule will include time for a peer review by a group of qualified and independent reviewers of select draft deliverables (so that review comments can be addressed when finalizing deliverables). The evaluator shall submit all project deliverables to the LBNL Project Manager and the DOE Project Manager.

Appendix B. Example of SOW for Non-R&D Evaluation Study

Evaluation Focus and Design

Subcontract Number ____

Evaluation Focus
The purpose of this evaluation is to assess the outcomes achieved by the DOE Wind Powering America (WPA) initiative through its state-based activities, and to understand the particular processes that proved most effective in achieving those outcomes. The evaluation will thus have both a process and impact element.

As summarized in the background of this Request for Proposal (RFP), the national strategy team established four thematic areas for organizing WPA initiative activities:

- State-based activities, centered primarily on the state wind working groups (WWGs);
- Rural economic development;
- Utility partnerships; and
- Federal wind power / Greening federal loads.

The focus of this evaluation will be *limited* to the state-based activities, with particular attention on the WWGs, who were recipients of targeted funding. In broad terms, this evaluation seeks to assess whether these targeted investments and activities were successful, and what aspects of the state-based activities enabled the success.

WPA activities in the areas of rural economic development, utility partnerships and the more generic national level support activities were to some extent integrated with state-based activities (see the background to this RFP). The evaluator's task is to focus on the investments and activities conducted under the rubric of state-based activities, and the outcomes attributable to those efforts. Thus part of the evaluation would be to identify the appropriate modalities for separating, as much as possible, the potential contaminating effects of the activities performed under the other thematic areas.

<u>Process Evaluation</u>
The process evaluation will focus on determining what program designs and implementation strategies by WPA's state-based activities resulted in: (a) the greatest success with respect to achieving the desired WPA outcomes and why; and (b) identifying what were less successful and the reasons why. The process evaluation of WPA's state-based efforts will include, but not necessarily be limited to, the following questions:

1. What elements of WPA's state-based activities have been most successful, and why?
2. Which WWGs have been most successful, and why? What are the characteristics of the successful WWGs (e.g., leadership, membership, direction, media participation, etc.), that fostered their effectiveness?
3. What, if any, common conditions were present for states where the WWGs were less effective (inherent market or environmental factors, level of complexity of projects)? What could be done to minimize these conditions in the future?
4. What are the lessons learned and best practices from this evaluation for use by WPA and DOE in light of its future plans for expansion of wind development across the U.S. (including both land-based and offshore wind)?

Impact Evaluation

The impact evaluation will address the following outcome metrics:

1. Megawatt (MW) capacity growth in states, before and after WPA, that is attributable to WPA.
2. Total amount of dollars leveraged by the states from DOE's investment for wind energy deployment activities.
3. Extent of replication that has occurred (e.g., best practices (are they doing something that they learned from WPA network?), information dissemination, new partnerships formed, new organizations emerging with similar mission, number of schools using Wind for Schools curriculum or erected a turbine, number of other stakeholder/partners who have erected a turbine, etc.).

Evaluation methods will be used to establish, as much as possible, the extent of influence of the state-based program activities on outcomes and to separate out rival explanations of other contributing influences. In addition to developing baselines against which to measure changes, the assessment should help determine an estimate of relative influence, taking into account what would have happened if WPA's activities had not occurred. Because of multiple activities by multiple players in the energy efficiency and renewable energy field, it will be challenging to isolate WPA's unique efforts in accelerating wind energy deployment. Nevertheless, the evaluator will develop plausible hypotheses, supported by counterfactual and attribution analysis, to quantify the impact of WPA's state-based efforts.

Evaluation Design

The WPA initiative was developed strategically to target key aspects of the wind energy market for maximum impact. The targeting meant, among other things, that not all states or stakeholders had a measureable and equal probability of participating in the initiative. For that reason, a randomized controlled trial (RCT) methodology cannot be applied to the current study. Instead, the impact evaluation will employ some form of a quasi-experimental, pre-post comparison study design. That is, the establishment of causality is usually best accomplished when pre- and post-measurements can be compared between an intervention group and a control group. The successful candidate would be expected to utilize an appropriate mix of evaluation methods – interviews, surveys, in-depth case studies, document reviews, etc. – to conduct the evaluation to establish some form of a credible quasi-experimental design.

The selection of appropriate methods should serve to address the internal validity issues associated with this study. One issue to be addressed is selection bias. For example, given the strategic targeting employed in the design of the program, a non-participating State may or may not represent a viable control, since its non-participation is specifically due to a selection "bias" that excluded that State from consideration, or subjected it to a much lower level of resource investment. Thus, the evaluator should consider alternative means for separating out participating and non-participating States to tease out the role WPA has played or could have played.

Another issue has to do with the nature of WPA's interventions. State-based activities, rural economic development, utility partnerships, and federal wind power/ greening federal loads are focused on a given State or a WWG, but are also accessible to everyone. This means that isolating one source of WPA information dissemination from another source might be

particularly difficult. Put differently, because many of WPA's various State activity outputs are purposefully broadcast and made available to everyone, it may be challenging to isolate the effects of one activity (e.g., State wind maps), from another medium (e.g., State conferences). The evaluator will need to identify some means of assessing the effect of interventions targeted to specific States (e.g., WPA States with WWGs) versus interventions targeted more broadly (all States receiving WPA resources).

TASKS (Refer to Appendix A)

Appendix C. Example of a Request for Proposal for a Program Evaluation Study[41]

The University of California Lawrence Berkeley National Laboratory
Procurement Department One Cyclotron Road Berkeley, CA 94720

Date:
Subject: Request for Proposals (RFP) No.:
Proposal Due Date:

The University of California, Lawrence Berkeley National Laboratory ("LBNL") requests a proposal for a retrospective benefit-cost evaluation of the U.S. Department of Energy's (DOE) nickel metal hydride (NiMH) battery research and development (R&D) investments, in accordance with this Request for Proposal (RFP) and the enclosed Proposal Preparation Instructions, the Sample Subcontract, and other enclosures.

Background
DOE's Vehicle Technologies Program (VTP) in the Office of Energy Efficiency and Renewable Energy (EERE) has sponsored R&D that has resulted in the state-of-the-art NiMH batteries used in all of today's hybrid electric vehicles (HEVs) (see References). This R&D began in the early 1970's.

The benefits from successful R&D investments typically accrue over lengthy periods, and may cut across different dimensions of society, thus requiring rigorous, well-designed evaluation studies to illuminate the impacts. For example, the identification of the role played by DOE in the development of NiMH batte1y technology emerged from an evaluation bibliometric study linking its funded R&D in advanced energy storage to a commercialized technology in today's hybrid, plug-in hybrids and electric vehicles. Using DOE's methodology for retrospective benefit-cost evaluation, this project will seek to extend the evidence to establish the economic, energy, environmental and energy security benefits of DOE's investment.

DOE has tasked LBNL to work with a subcontractor to perform an evaluation of DOE's NiMH battery research and development (R&D) investments.

Objective of the RFP
The objective of this RFP is to select a qualified subcontractor to estimate realized benefits and costs of the DOE's NiMH battery R&D investments using DOE's methodology for retrospective benefit-cost evaluation (see References below), as described in the Statement of Work. This methodology assesses the net economic benefits of cluster of specific technologies, as well as energy, environmental and security benefits.

It is expected that the work under the Subcontract will be performed over a period of 10 months. The anticipated not-to-exceed budget for the Subcontract is: $$$.

An original of a complete written (hard copy) proposal and two copies and an electronic version must be received by the undersigned LBNL Procurement Representative by 4 P.M. Pacific Standard Time on July 27, 2011. The proposal shall be valid for a period of 120 days from the proposal due date.

[41] This is example from a LBNL RFP used to hire a private evaluation firm via competitive solicitation. A different lab may have slightly different RFP requirements and format. EERE Golden Field Office also has its own RFP and RFQ requirements and format.

This solicitation is governed by procurement policies and procedures established under the Prime Contract between the University of California and the U.S. Government, represented by the Department of Energy ("DOE"), for management and operation of LBNL. Any award resulting from this solicitation will be a subcontract under the Prime Contract.

Acceptance of late proposals will be at LBNL's sole discretion. LBNL reserves the right to reject any and all proposals, to waive any minor irregularities in any proposal, or to cancel this RFP at any time prior to award without cost to LBNL. This RFP does not include provisions for the direct reimbursement of proposal preparation costs.

NAICS CODE AND SMALL BUSINESS SIZE STANDARD

The North American Industry Classification System (NAICS) Code for this acquisition is 541990, All Other Professional, Technical, Scientific Services. The corresponding small business size standard for this acquisition is annual receipts of $7 Million or less. Annual receipts are to be based on the average annual gross revenue for the past three fiscal years.

The Offeror shall complete the *Small Business Program Representations* clause in the attached Representations and Certifications form based on this small business size standard. Refer to Subpart 19.1 - Size Standards of the Federal Acquisition Regulation (FAR) for information on calculating the annual average gross revenue.

SUBMITTAL OF PROPOSALS

The Offeror shall send the proposal to the University Procurement Representative at the address shown below. Electronic submittals shall have the RFP# in the Subject line of the transmittal.

Contact:
Title:
Address:
Telephone No.:
Email:

BASIS FOR SELECTION – BEST VALUE

LBNL intends to select the responsive and responsible Offeror whose proposal contains the combination of supplier attributes and probable cost offering the best overall value. LBNL's selection may be made on the basis of the initial proposals or LBNL may elect to negotiate with Offerors selected as finalists.

A responsive Offeror is one whose offer satisfies the requirements of this RFP, including the technical requirements of the proposed subcontract. A responsible Offeror is one that is considered capable of performing and is otherwise eligible and qualified to perform the proposed subcontract.

LBNL will determine the best overall value by comparing differences in supplier attributes offered with differences in probable cost and related factors, striking the most advantageous balance between supplier attributes and the overall probable cost to LBNL.

LBNL reserves the right to make multiple awards as a result of this solicitation, if it is in the best interest of LBNL.

EVALUATION OF PROPOSALS

LBNL will evaluate each Offeror's proposal based on the information provided by the Offeror, LBNL's own experience, and/or information from the Offeror's customers. Offerors should, therefore, be persuasive in describing their supplier attributes and their value in enhancing the likelihood of successful performance and achievement of LBNL's objectives.

The evaluation of supplier attributes will focus on the strengths and weaknesses of the proposal within the framework of capability, affordability, and feasibility.

Capability

- What are the Offeror's supplier attributes (personnel, financial, etc.) and how well will they enable the Offeror to satisfy LBNL's requirements?
- What experience, knowledge and involvement in retrospective benefit-cost evaluation allows leverage of prior experience or organizational knowledge, assets and processes in this project? Note experience in conducting similar retrospective benefit-cost evaluations.
- What ability does the Offeror have to identify and utilize publicly available information about these energy programs and related activities?
- How has the Offeror demonstrated ability to write clearly and understandably, especially to a targeted audience?
- What is the Offeror's past experience on similar contracts? Has the Offeror successfully performed recent contracts similar in type and complexity as the proposed subcontract? The Offeror should include in its proposal a written description of recent contracts similar in type and complexity as this scope of work that the Offeror successfully completed. These may include public and private contracts. Include technical and business contact points by name, title, address, telephone number and, if available, e-mail address. Offerors are encouraged to include a self-assessment of their performance on these contracts, including what went well and what did not. Offerors may discuss the latter in the context of a lessons learned from the project and how the lessons have been incorporated into improved products/services and processes.
- If work will be subcontracted, to what extent is the assignment of work scope appropriate, and to what extent is the prospective lower-tier subcontractor qualified to do that work?

Affordability

- What is the probable cost to LBNL?
- How realistic is the proposed estimate/budget in relation to the approach that the Offeror proposes to employ?
- How does the proposed estimate/budget compare to the LBNL's estimate and to other proposed budgets, and are the trade-offs worth the difference?

Feasibility

- How well will the proposed approach contribute to successful and timely completion of the work?
- How well will the approach work as proposed?

- What is the probability that the proposed study will be provided within the time frame required?

ENCLOSURES

1. The Offeror shall complete the following enclosure and submit it with the proposal:
    a. Representations & Certifications
    b. OCI Disclosure
    c. Statement Budget
    d. Proposal
2. Pre-award Survey of Prospective Subcontractor's Accounting System

The following enclosure will be required only from the successful Offeror:

3. Offeror: E-verify Registration

PROPOSAL PREPARATION INSTRUCTIONS

PROPOSAL CONTENTS

General

The proposal should consist of a technical/management proposal and a cost proposal. The proposal should be submitted with a cover letter identifying the Offeror's name and address, solicitation number and title, the name(s), title(s), and telephone number(s) of the individuals in Offeror's organization who have commitment authority on behalf of the Offeror and will be responsible for contractual negotiations and administration of any resultant Subcontract.

Technical/Management Proposal

The technical/management proposal should contain a comprehensive discussion of Offeror's approach for successful performance of the work, plus any other element that may be necessary to fulfill the requirements of the Scope of Work, including a delivery schedule, and task plan with budget. LBNL requires delivery of several reports by 10 months after the start of the contract. Alternate delivery date may be proposed which may be subject to negotiation prior to award.

Supplier Attributes

The proposal should identify, describe, and discuss the supplier attributes that the Offeror considers important to successful performance of the proposed subcontract. LBNL has identified the supplier attributes listed in below, which are the qualitative criteria that the University will use for the subjective evaluation of proposals. They are not listed in any order of importance and no attribute is more important than any other. The Offeror should discuss them in the proposal and may identify other supplier attributes that it believes may be of value to LBNL. If LBNL agrees, they will be considered in the evaluation process. In all cases, LBNL will assess the value of each proposal as submitted.

Cost Proposal From Educational or Non-Profit Institutions

The cost proposal must include a total estimated cost for the work. In order to help establish cost realism, the estimate shall be supported by the following information:

- The hourly direct labor rate(s), the proposed hours, and the extended cost for each labor category that will be used in performing work under the resulting Subcontract. Specify rates and escalation factors used.
- Total proposed direct labor cost.
- Fringe benefit rate(s) (if applicable), extended to total fringe benefit cost.
- Overhead rate(s) - full rate and applicable base, extended to total overhead cost. Identify that portion of the rate attributed to depreciation of facilities capital. (Example: Overhead is I00 percent of direct labor dollars; 4.3 percent is attributed to depreciation).
- General and Administrative Expense rate(s) - full rate and applicable base, extended to total G&A cost. Identify the portion of the rate attributed to depreciation of facilities capital and to independent research and development. (Example: G&A is 11.5 percent of total direct costs and overhead; 2 percent is attributed to depreciation, 1 percent to R&D).
- Travel costs (itemize); domestic only.
- Materials, subcontracts and services (itemize).
- Other direct costs (itemize).
- Total Estimated Cost.
- Facilities Capital Cost of Money (if applicable). Identify applicable rates and bases, extended to total FCCM cost.

The Offeror shall provide copies of current forward pricing or other rate agreements reached with a cognizant Government agency if those rates are used in the proposal. LBNL reserves the right to examine, at any time prior to award, any of those books, records, documents, or other records directly pertinent to the information requested or submitted. Depending on the circumstances, Offerors may be required to submit pricing information as defined in the Federal Acquisition Regulation (FAR) 15.4, including certification of the data as current, complete, and accurate upon conclusion of negotiations.

Financial Capability

The cost/price proposal should fully describe the Offeror's current financial condition and its financial ability to support the performance of the proposed subcontract. Upon request, the Offeror shall provide financial information such as the following:

1. Audited and certified year- end financial statements for the last two years (balance sheet, income statement and other financial statements or reports);
2. Financial statements compiled and reviewed by a certified public accountant or other accounting professional (include the accounting firm's cover letter);
3. Tax returns for the two most recent completed fiscal years; or
4. Other information acceptable to LBNL. LBNL reserves the right to request additional financial statements.

Organizational Conflicts of Interest (OCI) Disclosure

The selected Offeror shall provide an OCI Disclosure Statement, utilizing the attached form. Organizational conflict of interest means that because of other activities or relationships with other persons, a person is unable or potentially unable to render impartial assistance or advice to the Government, or the person's objectivity in performing the subcontract work is or might be otherwise impaired, or a person has an unfair competitive advantage.

The Disclosure Statement must contain the following:

1. A statement of any past (within the past twelve months), present, or currently planned financial, contractual, organizational, or other interests relating to the performance of the statement of work.
   a. For contractual interests, such statement must include the name, address, telephone number of the client or client(s), a description of the services rendered to the previous client(s), and the name of a responsible officer or employee of the Offeror who is knowledgeable about the services rendered to each client, if, in the 12 months preceding the date of the statement, services were rendered to LBNL, the University, or any other client (including a U.S. or foreign government or person) respecting the same subject matter of the instant solicitation, or directly relating to such subject matter. The LBNL or University contract number under which the services were rendered must also be included, if applicable.
   b. For financial interests, the statement must include the nature and extent of the interest and any entity or entities involved in the financial relationship. For these and any other interests, enough such information must be provided to allow a meaningful evaluation of the potential effect of the interest on the performance of the statement of work.
2. A statement that no actual or potential conflict of interest or unfair competitive advantage exists with respect to the advisory and assistance services to be provided in connection with the instant subcontract or that any actual or potential conflict of interest or unfair competitive advantage that does or may exist with respect to the subcontract in question has been communicated as pai1of the required statement.

Failure of the Offeror to provide the required statement may result in the Offeror being determined ineligible for award. Misrepresentation or failure to report any fact may result in the assessment of penalties associated with false statements or such other provisions provided for by law or regulation.

Offeror's Questions

LBNL will respond to questions submitted in writing to the LBNL Procurement Representative on or before July 15, 2011. Questions submitted after this date may not be answered prior to the proposal due date. Questions may be submitted by letter, facsimile or e-mail, with e-mail preferred. Answers to questions that are germane to the interpretation of LBNL' requirements will be issued to all Offerors in writing.

Acceptance of Terms and Conditions

Submission of a proposal shall indicate the Offeror's willingness to accept the terms and conditions of the Sample Subcontract and its attachments unless specific exceptions are taken. These terms and conditions have been approved by the DOE. Failure to accept the terms and conditions will not be evaluated favorably, and may cause LBNL to reject Offeror's proposal.

Proprietary Information

LBNL will treat any commercial or financial information in the proposal as proprietary. LBNL prefers not to receive proprietary technical information. If the proposal includes any proprietary technical information, it must be marked "Proprietary" or its equivalent. LBNL

will use its best efforts to (I) maintain such proprietary information in confidence, giving it the same degree of care, but no less than a reasonable degree of care, as LBNL exercises with its own proprietary information to prevent its unauthorized disclosure; and (2) only disclose such proprietary information to its employees, agents, consultants, subcontractors or Government personnel who have a need to know related to this RFP and are bound by an obligation of confidentiality.

If the Offeror intends to use a product or process in which there is a proprietary or background patent position, the proposal should indicate and list patent applications and/or patents granted (including dates, numbers, and descriptions), and whether the Government has rights to the patents.

PROPOSAL FORMAT

Offerors are requested to provide concise yet complete description of the Offeror's approach and capabilities for satisfying the required services outlined in this RFP. Excessive length is discouraged. In addition, Offerors are encouraged to pro-actively present additional information and responses, not specifically requested, that help demonstrate understanding of LBNL's specific evaluation objectives and needs as well as bidder creativity, experience, and/or expertise.

Proposals must adhere to the following set format (the numbers indicated are suggested page limits):

- Proposal cover
- Signed cover/transmittal letter
- Table of Contents (include proposal date and page numbers on each page of proposal)

Sections

1. Company overview
2. Executive summary (2 pages)
3. Work scope and schedule (15 pages)
4. Responses to essay questions (4 pages)
5. Staffing and subcontracting (5 pages)
6. Qualifications and Experience (10 pages)
7. Budget (2 pages plus tables)
8. Disclosures and required documents (as needed)
9. Appendix – Resumes (2-pages per resume)

Proposal Cover and transmittal letter

The proposal cover should indicate the RFP name, the proposal date, Offeror's name, and list of subcontractors with an indication of which portfolio is addressed in the proposal. The cover should also state that the person signing the letter is authorized to commit Offeror's organization to the proposed work scope, budget and rates, terms and conditions of the RFP, and that the information in the proposal is accurate.

Sections 1and 2: Company Overview and Executive Summary

Sections 1 and 2 of the proposal contain general information about Offeror's firm, and a high level summary of the proposal including the approach to LBNL's evaluation tasks and the bidding team's qualifications to perform the services sought through this RFP.

Section 3: Work Scope and Schedule

Section 3 of the proposal should discuss Offeror's approach to conducting a retrospective benefit-cost evaluation for the work described in the Statement of Work and the Project Description documents. Offerors are reminded that an initial task for the selected consultants will be to develop a detailed evaluation plan. Section 3 of the proposal should describe Offeror's approaches with sufficient detail to distinguish the strengths and unique features that are suggested, but it should not be overly detailed and lengthy.

Section 3 should include a schedule for performing Tasks. The schedule should be presented graphically and supplemented with text explanations needed to provide a complete understanding of the proposed timeline. The schedule will be reviewed at the project's kickoff meeting.

Section 4: Responses to Essay Questions

To encourage Offerors to demonstrate their creativity and their understanding of LBNL's requirements, Offerors are required to include short essay responses to the three (3) questions posed below. The essays should be concise yet comprehensive, and they should reflect an understanding of evaluation best practices and LBNL's needs.

1. LBNL is interested in the effects of possibly increasing or decreasing the evaluation budget. What would be the impact if the evaluation budget was reduced by 25%? What additional services and/or improvements are suggested if the budget was increased by 25%?
2. Discuss what kinds of issues should be considered when attempting to use the vast existing program evaluation literature and analyses to provide recommendations to LBNL on DOE's NiMH battery R&D investments. What kinds of information might map well for achieving LBNL's goals, and how might existing information be supplemented to reflect those circumstances?
3. Describe how Offeror would suggest handling accuracy and uncertainty in measuring and reporting program results? What strategies will work well for reducing overall uncertainty while controlling evaluation costs? How does Offeror propose treating and reporting uncertainty of the results?

Section 5: Staffing and subcontracting plan

In this section, Offerors are requested to:
Include a management and organizational chart that depicts the relationships among team members to accomplish the proposed work. Note that LBNL expects that the primary contact with LBNL will be the Offeror's project manager.

Identify the lead staff member assigned to manage the evaluation work, provide a short biography, and explain why he or she is qualified for this position. Describe this person's availability for the project, and the office where he or she will be based. Identify the key personnel to be assigned to this project, describe their responsibilities. Indicate availability and length of time commitment to project. Include resumes for all individuals named in budget in a proposal appendix. Resumes and bios should describe relevant responsibilities from other projects that will help LBNL evaluate the qualifications and experience of key personnel.

Please limit length of resumes to **two** pages. Staffing changes for key personnel are subject to approval by LBNL.

Section 6: Qualifications and Experience

Use this section to address Offeror's qualifications and experience, drawing on lessons learned and personal best practices. At a minimum, please address the following topics, although Offerors are not limited to these alone:

6.1 Summary

Summarize why Offeror is best suited to conduct the requested Evaluation, Measurement & Verification (EM&V) services. Include any previous experience that the bidding team has in evaluating utility and government R&D programs.

6.2 EM&V Experience

Describe Offeror's previous experience as the lead consultant responsible for a retrospective cost-benefit evaluation.

List recent, last three years, evaluation reports and peer-reviewed literature prepared by team members that are relevant to this RFP's scope and programs.

Describe Offeror's experience in successfully completing evaluation assignments. Include at least the following descriptions of team experience with:

- Quantitative estimation of economic benefits and costs of technologies.
- Computation of energy, environment and energy security performance measures of technologies.
- Technology and market assessments.
- Data collection as it relates to the proposed study.
- Choice of the counterfactual, including a next best alternative to a technology.
- Incorporation of bibliometric data into analysis of benefits and attribution.
- Assessment of attribution (e.g., "isolating" the effects of DOE R&D investments from other possible influences on the benefits and costs).
- Uncertainty analysis for retrospective evaluation. Developing and reporting uncertainty levels for evaluating benefits and costs of R&D investments.
- Describe any innovative methods that you may have used to mitigate uncertainty and bias in your reporting of the benefits and costs.
- Presenting evaluation findings to diverse audiences.
- Managing teams of similar size for similar evaluation activities.

6.3 References

Offerors should provide three or four references from current (preferred) or recent clients for whom they have performed projects that are relevant to the work scope. References should include a brief synopsis of specific services provided, company name and location, contact name, contact title, telephone number, and email address of the reference.

Section 7: Budget

Using the forms provided with this RFP package as a separate Excel file, Offeror should provide labor and other direct costs for evaluation services to be provided during the period

of this contract. Provide Tables 7-1 and 7-2. Examples of completed tables are included in the Excel Budget Form. These are for demonstration of the use of the tables, and they are not indicative of budgets, budget allocations, staffing categories or any other aspect of the budget.

The completed Excel budget file should be submitted as a separate document. Please include Offeror's Name in the file name.
The budget tables should be "pasted" into the proposal document as part of Section 7.

Time and Materials cost proposals from Offerors should include the following information:

- For each proposed labor category, the fixed fully burdened hourly labor rates (specify the applicable time period and any escalation factors used), the proposed hours, the extended cost, and the total proposed direct labor cost.
- Travel costs (showing estimates for air and ground travel and miscellaneous and incidental expenses (M&IE) and lodging estimates based on CONUS per diems, itemized by number of travelers, destinations, and trips).
- Materials, subcontracts, and services (itemized and stating applicable rates and the basis for/source of the estimates).
- Total proposed non-labor cost.
- Total estimated cost.
- A budget corresponding to the Task Plan required for the technical/management proposal should also be included.
- A completed copy of the enclosed Pre-award Survey of Prospective Subcontractor's Accounting System.

Section 8: Disclosures and Documentation (as needed)

Section 9: Resumes

(End of Proposal Preparation Instructions)

References
Representations & Certifications
General Provisions for Cost Reimbursable (no fee) Subcontracts

Appendix D. Procedures for Obtaining OMB Approval to Collect Information

The Paperwork Reduction Act (PRA) of 1995[42] requires that every federal agency obtain approval from the Office of Management and Budget (OMB) before undertaking to collect information from ten or more persons, or before continuing a collection for which the OMB approval and the OMB control number are about to expire. The approval process, which is popularly known as the "OMB clearance process," can be extensive and time-consuming. Usually, it requires two Federal Register notices and a detailed application to OMB called an "Information Collection Request." The duration for the entire process can exceed six months.

Four exceptions exist to this lengthy process for obtaining OMB clearance to collect information. These exceptions are: (1) surveys of substantially the same intent that must be conducted on short notice to satisfy public policy needs; (2) surveys that measure participant satisfaction with a program; (3) emergency reviews when a data-collection activity must be performed to meet an unanticipated, urgent need and no time is available for public comment; and (4) surveys of Federal Government employees on subjects concerning their employment. Only the first two of these exceptions have potential relevance for program evaluations.

The process for earning OMB approval of the first two of these exceptions is called a "generic clearance" process. Any evaluation-related data collection that does not meet the requirements for a generic clearance application is called the "full clearance" process. It will require the Federal Register notices and detailed Information Collection Request (ICR).

Generic Clearance Process

The first exception to the full clearance process exists to approve data collections that will measure customer satisfaction with a program. The second type of information collection may be used to test evaluation data-collection methods such as a data-collection process, a questionnaire, or testing a message.[43]

If you think your data collection activity matches one of these generic clearance requirements, contact the DOE Office of the Chief Information Officer (OCIO), explain your planned data collection and request guidance on how to proceed.[44] If OCIO judges that the data collected by your collection plan will be analyzed statistically, you will be instructed on how to submit your application for a generic clearance to the Energy information Administration (EIA) for further review and submission to OMB. If the data produced by the proposed data collection will not be analyzed using statistical procedures, e.g., information from a focus group of ten or more persons, the request for a generic clearance will be submitted to OMB by OCIO, and you will be advised on how to submit your application to OCIO. OCIO will advise you on the time expected for OMB's approval; it can range from two weeks to several months.

---

[42] 44 U.S.C. chapter 35; 5 CFR Part 1320.

[43] The types of information collections that may be approved using DOE's generic clearance are described in more detail in an OMB memorandum on Generic Clearances dated May 28, 2010 that is available at http://www.whitehouse.gov/sites/default/files/omb/assets/inforeg/PRA_Gen_ICRs_5-28-2010.pdf (last accessed June 17, 2014).

[44] OCIO's point of contact (June 11, 2013) is Chris Rouleau, PRA Officer. 301.903.6227. Christina.rouleau@hq.doe.gov.

The Full Clearance Process

All other types of information collections in support of evaluations must be approved through the full PRA clearance process. This process is managed for DOE by OCIO.

OMB offers a one-time full-clearance process option for program evaluations or other studies that will collect substantially the same information at different times during a three-year period. OMB calls this a "multi-stage clearance."[45] It avoids the necessity of having to go through the full PRA clearance process for the subsequent information collections after the first one. Contact OCIO for guidance on when this type of full clearance is appropriate.

Time for OMB to Complete the PRA Approval Process

DOE has its own internal requirements for a PRA submission; therefore, the program should expect to work closely with the appropriate DOE or EIA office that will be its point of contact with OMB. This is very important.

Although OMB reports that the median elapsed time for its review after it receives the information collection request from the agency is 62 days,[46] in the case of evaluations this process may require six months or more. This time may be reduced for an information collection request that is well thought-out and justified.

Process Diagrams

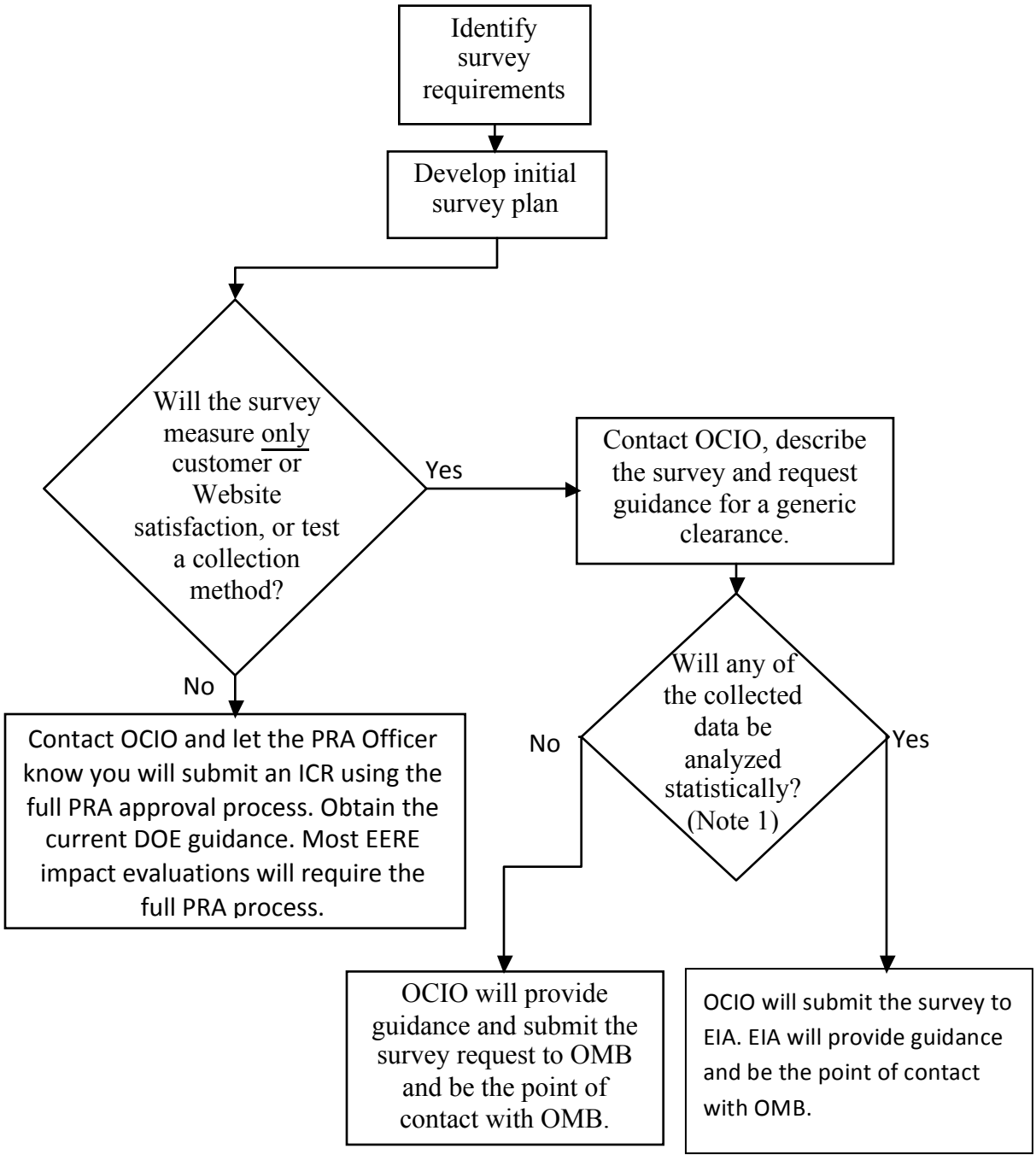
This appendix includes the following process flow charts to help you navigate the OMB clearance process:

- Determining Which OMB Clearance process to Use
- The Generic Clearance Process
- The Full PRA Clearance Process.

The flow charts show the basic decision points for selecting the appropriate clearance process and the steps for obtaining clearances under the two processes. Notes to the charts provide additional guidance. The charts use the word "survey" to refer to all information collections although some may use other methods, e.g., focus groups.

[45] OMB presentation, "Dispelling Myths About The Paperwork Reduction Act (PRA)," Presented to a meeting of the Federal Evaluators' Network. March 13, 2013. Not available online.
[46] Ibid.

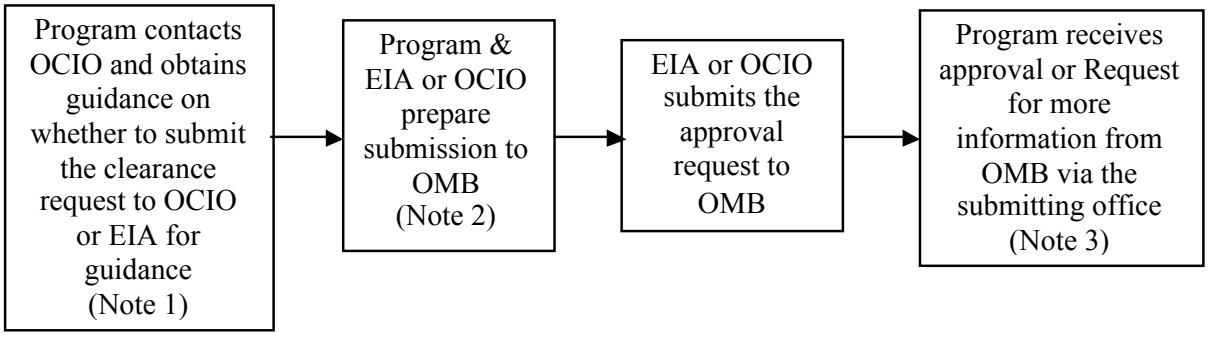
Determining Which OMB Clearance Process to Use



Note:

1. Statistical analysis involves (a) estimating precision intervals for an estimate obtained by a random survey, and/or (b) a hypothesis test regarding whether the estimated difference between two measurements might have been due to chance.)

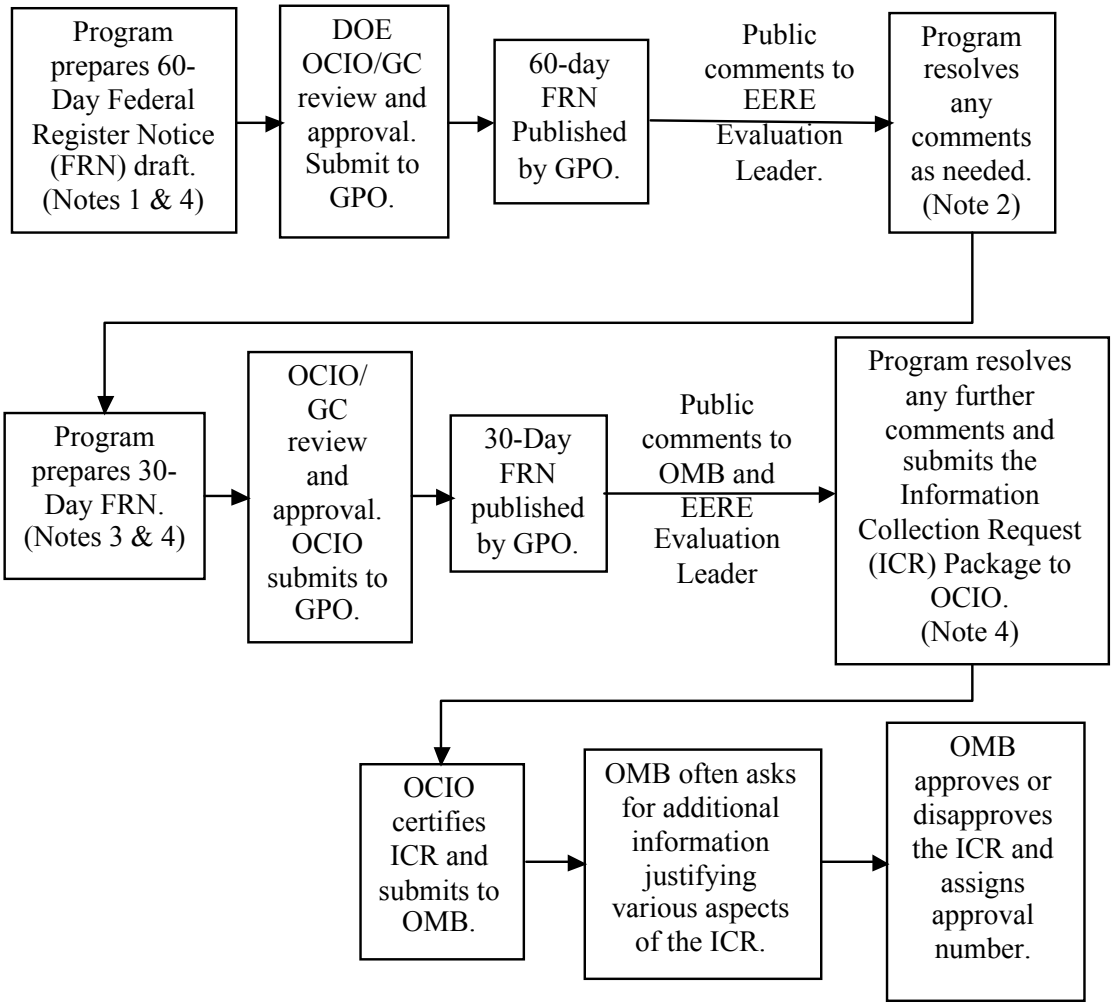
The Generic Clearance Process



Notes to the Generic Approval Process Diagram:

1. Contact DOE Paperwork Reduction Act Officer at OCIO's Records Management Division at 301-903-6227. If OCIO decides that the data will be collected for statistical analysis, EIA will guide the submission and be the point of contact with OMB. Otherwise, OCIO will perform these functions for the program.
2. Information to include in the letter to OMB requesting generic clearance:
   - Identify organization(s) that will be using the clearance
   - Reference generic clearance agreement (OCIO or EIA will provide this)
   - Make commitment to provide survey results to OMB
   - The data collection instrument (termed the "protocol").
   - Information about the survey the program can expect to provide to EIA or OCIO for the request. (These requirements change from time to time; check with the respective office for the current requirements.) For example:
     - The function(s) and objective(s) of requesting organization(s)
     - Reasons for developing the survey
     - How survey responses will be used (e.g., improved program delivery)
     - Description of survey respondents (what population will be surveyed?)
     - Survey distribution and response collection method
     - Estimate of response rate/number of respondents
     - Participation factors (type of respondent, voluntary participation, data confidentiality)
     - Estimate of time burden for responding for a single respondent and the total estimated burden (single response time multiplied by the estimated number of responses)
     - Whether data will be statistically analyzed
     - Types of statistical analyses to be used (for EIA only).
3. OMB has targeted two weeks to approve a generic request or request more information.

The "Full" PRA Clearance Process



Notes to the Full PRA Approval Process Diagram:

1. Obtain the template for the 60-day Federal Register notice (FRN) from the DOE OCIO's Paperwork Reduction Act Officer (301-903-6227) and submit your draft to that office. OCIO will handle DOE's General Counsel review and submission to the Government Printing Office (GPO).

The FRN for the collection should include the following information:

- A statement that DOE is proposing an information collection to OMB
- The title of the information collection
- A summary of the information collection
- A brief description of the need for the information and its proposed use
- (The notes to the Full PRA Approval Process diagram continue on the next page)

Notes to the Full PRA Process (Continued)

A description of the likely respondents and proposed frequency of response
A summary of the privacy risks involved in collecting the information electronically from potential respondents (if appropriate)
An estimate of the total annual reporting and record keeping burden
Direction that comments should be submitted to DOE (see below).
The FRN should indicate that public comments and requests for supporting information, e.g., the draft ICR (see Note 4), should be submitted to DOE within 60 days of the date of publication in the Federal Register. The notice must indicate where the comments can be sent. This may be EERE's Evaluation Lead or the sponsoring program office.

2. If comments are received during the 60-day public comment period, they must be evaluated and responses must be prepared. A summary of the public comments, including those actions taken in response to the comments, must be included in the ICR package (Note 4). In particular, changes to the proposed information collection that result from the public comments must be described.

3. After the 60-day FRN has been published in the Federal Register and appropriate action has been taken to address any public comments received, a 30-day FRN must be prepared to notify the public that the information collection is being submitted to OMB for review and approval. DOE OCIO may ask for the proposed ICR at this point.

   Obtain the template for the 30-day FRN from the OCIO office listed under Note 1. The 30-day FRN should include the following information:

   - A statement that OMB approval is being sought
   - The title of the information collection
   - A summary of the information collection
   - A brief description of the need for the information and its proposed use
   - A description of the likely respondents and proposed frequency of response
   - An estimate of the total annual reporting and record keeping burden
   - Indication that comments be submitted to OMB
   - Statutory authority for collecting the information.

4. Work to develop the draft ICR should begin when the 60-day FRN notice is being prepared. The FRN will include information that is part of the ICR. The complete ICR package, including responses to public comments from the 60-day FRN, should be submitted to the OCIO along with the 30-day FRN.

   OCIO's Paperwork Reduction Act Officer can provide guidance for completion of the ICR package. The package will consist of the following documents:

- OMB Form 83-I, "Paperwork Reduction Act Submission"
- The Supporting Statement, Part A and if the information collection is designed to be analyzed statistically, Part B.
- Draft information collection tools/instruments (e.g., forms and accompanying instructions, copy of the citation from the governing regulation)
- Summary of public comments received
- Supporting documentation in accordance with OMB's current requirements.

5. OMB Form 83-I requires certification that the proposed collection of information:
    - Is necessary for the sponsoring program office to perform its functions
    - Avoids unnecessary duplication
    - Reduces the burden on the respondents
    - Uses plain, coherent, and unambiguous terminology
    - Will be consistent and compatible with current reporting and record-keeping practices
    - Indicates the retention period for record-keeping requirements
    - Informs respondents about:
        - Why the information is being collected
        - How the information will be collected
        - How the information will be used
        - The extent of the estimated labor and cost burden to respond
        - The nature of response expected (voluntary, required, or mandatory
        - The level of confidentiality to be imposed
        - The requirement that a valid OMB control number must be displayed.
        - Was developed by an office that has planned and allocated resources for managing the information to be collected
        - Does or does not use statistical survey methodology
        - Does or does not use information technology to reduce burden.

    The Supporting Statement Part A documents that these certification requirements have been met. If the information collection will be based on statistical sampling methodology, Part B of the Supporting Statement must also be completed and submitted.

    Evaluation contractors are likely to ask whether an incentive can be offered to encourage a respondent to take a survey. Part A of the ICR Supporting Statement requires the program to state whether it intends to offer any payment of gift to respondents. OMB does not offer "official" guidance on this question. Unofficially, OMB staff have said that OMB will approve incentives only for information collections it judges to have high burden (require an interview of more than 15 minutes) or to be invasive (probe potentially sensitive personal actions or opinions), but it will require that the program submit evidence of the effectiveness of incentives for similar surveys. The program should consult OCIO on the use of incentives for focus groups.

Sources for Further Information on the OMB Clearance Process

DOE's PRA Officer has prepared a training presentation for program managers who will collect information from the public. This 23-page document is a good overview of the OMB clearance requirements and DOE's process for using it. This document can be downloaded from: http://energy.gov/sites/prod/files/cioprod/documents/Information_Collection_Program_Training.pdf (last accessed June 2014).

The Executive Office of the President has published a more comprehensive guide that provides OMB's requirements. It may be downloaded from: http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/pmc_survey_guidance_2006.pdf (last accessed June 2014).

Appendix E. Example of a Non-R&D Evaluation Report Outline

Draft Final Report

1. **Executive Summary**
   a. Evaluation Objectives
   b. Methodology
   c. Findings
   d. Key Lessons Learned
   e. Recommendations

2. **Introduction**
   a. Program Description
      i. Program Terminology
      ii. Summary Reported Program Accomplishments
   b. Evaluation Goals and Objectives
   c. Summary of Key Evaluation Activities
   d. Evaluation Challenges
      i. Difficulty Interpreting Grantee Data
      ii. Inaccuracies of DOE Reported Metrics
      iii. Delayed or Lack of Grantee Responsiveness
      iv. Limited Value of Participant Phone Verification Surveys
      v. Large Scope and Broad Scale of Grantee Programs
   e. Report Outline

3. **Better Buildings Neighborhood Program (BBNP)**
   a. BBNP Goals, Objectives and Expected Program Effects
   b. Program Requirements
   c. Grantee Program Details
      i. Technologies and Services
      ii. Financial Incentives
   d. Reported Program Accomplishments
   e. Databases and Data Tracking Processes
      i. Grantee Data
      ii. DOE Reporting Processes

4. **Methodology**
   a. Obtain DOE Program Records
   b. Develop the Sample Approach
      i. Develop Specific Evaluation Activity Sample Frame
   c. Design the M&V Sample
      i. Determine the M&V Sample Parameters

ii. Establish the Sample Size
iii. Stratification

d. Conduct Measurement and Verification
    i. Obtaining Grantee Project Records
    ii. Designing the Survey and Data Collection Instruments
    iii. Conducting Telephone Verification Surveys
    iv. Conducting On-site Verifications
    v. Conducting Project File Reviews
    vi. Establishing the Baseline Scenarios
    vii. Verifying Gross Impacts
e. Conduct Billing Regression Analysis
    i. Data Cleaning
f. Review of Independent Evaluation
g. Net-to-Gross Methodology
h. Extrapolation of Results to Overall BBNP
    i. M&V Sample Extrapolation
    ii. Billing Analysis Extrapolation
    iii. Overall BBNP Extrapolation
i. Calculation of Additional Metrics
    i. Lifetime Energy Savings
    ii. Greenhouse Gas Emission Savings
    iii. Cost Savings
    iv. Demand Savings
j. Economic Impacts Analysis
    i. Analysis Methods
    ii. Model Input Data

5. Findings

a. Overall Preliminary Evaluation Findings
b. Measurement and Verification Findings
    i. M&V Sample Extrapolation
    ii. Issues that Impacted the M&V Findings
c. Billing Analysis Findings
d. Combining the M&V and Billing Analysis Findings and Extrapolating to the Population
e. Risks to Validity of Findings for the Approach Used
f. Confidence & Precision
g. Net-to-Gross Analysis Findings
h. Economic Analysis Findings

i. Gross Economic and Fiscal Impacts
ii. Net Economic Impacts
iii. Post-BBNP Energy Savings Impacts

6. Lessons Learned, Recommendations and Conclusions

a. Lessons Learned
   i. Grantee Interaction
   ii. Sampling
   iii. Evaluation Activities
   iv. Department of Energy

b. Recommendations
   i. Recommendations for the Final Evaluation
   ii. Short-Term Recommendations for DOE
   iii. Long-Term Recommendations for DOE

c. Conclusions

7. Appendices

8. Figures

9. Tables

Appendix F. Example of an R&D Evaluation Report Outline

Draft Final Report

1. **Executive Summary**
   a. Introduction
   b. Methodology
   c. Findings
      i. Energy and Resource Benefits
      ii. Environmental Benefits
      iii. Energy Security Benefits
      iv. Knowledge Benefits
      v. Overall Economic Performance Measures
   d. Sensitivity Analysis
   e. Conservative Nature of Reported Results

2. **Introduction**

3. **VTO R&D Investments in Energy Storage Technologies**
   a. Genesis of DOE Funding for Electric Drive Vehicles
   b. Cumulative VTO R&D Investments in Energy Storage Technology
   c. VTO R&D Investments through the U.S. Advanced Battery Consortium

4. **Market Adoption of NiMH and Li-ion-Powered Vehicles**

5. **Methodology**
   a. Benefit-Cost Analysis Framework
   b. Conceptual Approach to Benefit-Cost Analysis
   c. Approach to Energy and Resource Benefits Estimation
      i. Primary Data Collection Protocol
      ii. Sample of Interviewees and Respondents
      iii. Estimation of Benefits Relative to the Next Best Alternative
      iv. Attribution of Benefits to VTO's Investments in NiMH and Li-ion Battery Technologies
   d. Approach to Environmental Impacts Estimation
   e. Approach to Energy Security Impacts Estimation
   f. Approach to Knowledge Impacts Estimation
   g. Measures of Social Economic Return

6. **Findings**
   a. Energy and Resource Impacts
      i. Counterfactual Battery Life, Energy Density, and Cost Improvement without VTO's R&D Investments
      ii. Counterfactual EDV Adoption without VTO's R&D Investments
      iii. Fuel Savings from VTO's R&D Investments
   b. Environmental and Energy Security Impacts
      i. Avoided Greenhouse Gas Emissions
      ii. Avoided Non-GHG Air Pollutant Emissions
      iii. Environmental Health Benefits

Appendix G. Example of an Evaluation Study Peer Review Charter

Introduction / Background on EERE and Program to be Evaluated

The mission of the U.S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE) is to strengthen America's energy security, environmental quality, and economic vitality in public-private partnerships that:

- Enhance energy efficiency and productivity
- Bring clean, reliable and affordable energy technologies to the marketplace, and
- Make a difference in the everyday lives of Americans by enhancing their energy choices and their quality of life.

To achieve this mission, EERE funds $1.2 billion of research, technology development, and demonstration and deployment programs. As a Federal office, EERE's role is to invest in high-risk, high-value research and development that is both critical to the Nation's energy future and would not be sufficiently conducted by the private sector acting on its own. EERE also works with stakeholders to develop programs and policies to facilitate the deployment of energy efficient and clean energy technologies and practices. For more information see the office website at http://www.eere.energy.gov/.

Within EERE, the **Weatherization Assistance Program (WAP),** created by Congress in 1976 under Title IV of the Energy Conservation and Production Act, functions "to increase the energy efficiency of dwellings owned or occupied by low-income persons, reduce their total residential expenditures, and improve their health and safety, especially low-income persons who are particularly vulnerable, such as the elderly, persons with disabilities, families with children, high residential energy users, and households with high energy burden" (Code of Federal Regulations, 2005).

In the early 1990's, as a way to provide policy makers and program implementers with up-to-date and reliable information for effective decision-making and cost-effective implementation, DOE sponsored a comprehensive evaluation of WAP. Subsequent years saw an increasing reliance on meta-evaluations to properly and accurately determine the benefits of the Program– an approach that has proven insufficient to the task. The current Evaluation Plan uses a new research design and takes into account the full range of substantive changes that have taken place both within the Program as well as contextually, in the years since the last comprehensive evaluation. This includes the incorporation of new funding sources, management principles, audit procedures, and energy-efficiency measures motivated by the findings and recommendations of the previous national evaluation, as well as the Weatherization *Plus* strategic planning process and other federal, state and local initiatives.

Main Objectives of the WAP Evaluation

As noted in the draft "National Evaluation of the Weatherization Assistance Program: Evaluation Plan for Program Year 2006," the purposes of the new evaluation are three-fold, notably:

1. To provide a comprehensive review of Program performance;
2. To enable DOE to make any necessary improvements and guide the direction of the Program into the next decade;
3. To promote the Program and further leveraging.

To address the stated purposes, the Evaluation Plan Design team proposed a series of studies that are appropriately summarized as: a) basic data collection studies, b) impact assessment studies, c) programmatic process assessment studies and d) process assessment studies for special measures and activities. The current review is intended to focus on the first two types of studies.

Purpose of the Expert Review of the WAP Evaluation

The WAP Evaluation includes an impact assessment (Section 2) as well as a process assessment element (Section 3). These assessments are to be completed over 4-year period.

The primary purpose of the Review is for the WAP manager and the Evaluation Design Team to obtain external review of the work plan, data collection approaches and instruments, data analysis approaches and analysis results for the 2006 WAP Evaluation Study. The Review will include both short and long term involvement by an expert panel of evaluation peers. The short term will focus on review of the WAP Evaluation Plan by a panel of experts. The long term will involve review of the draft WAP Evaluation Report by a panel comprised of many of the same experts (to ensure continuity in the review).

The current Review Charter is focused on the WAP Evaluation Plan and the impact assessment portion of the Plan. A second separate review process will be established to review the process elements of the Plan.

The WAP seeks the informed opinions and advice of external experts on the impact assessment portion of the draft WAP Evaluation Plan to improve the planned evaluation before implementation, thus lending substantive credibility to the results of the program by increasing the likelihood that the findings can withstand the most rigorous of examinations.

The primary audience for the Review is the WAP Manager and the Evaluation Design Team. The design team is comprised of Oak Ridge National Laboratory (ORNL) experts with a broad range of experience with the Program and its attendant subject area requirements. In their effort, the design team received input from a variety of sources, including WAP staff, the Network Planning Committee, EERE evaluation staff and other external stakeholders, and reviewed a variety of internal and external sources.

Review Chairperson and Panel

A panel of six experts lead by the Review Chairperson will be assembled for this review. The Panel will review the impact assessment portion of the WAP Evaluation Plan. Approximately twelve to eighteen months later the Chairperson and members of the panel will be reconvened to provide a review of the draft WAP Evaluation Report.

Role of Chairperson

The Chairperson of the review is an objective, unbiased, and independent expert from outside the program being reviewed. The review chairperson serves a unique and important role that can begin as early in the review process as he/she is selected, including the selection of the members of the peer review panel. Areas where the chairperson provides direction, oversight, and possibly final decisions can include the following:

- Selection of reviewers.
- Establishing review criteria.
- Establishing the content and scope of material submitted by research teams.
- Ensuring independence of the panel members during the review and the independence of the review more generally.
- Facilitating the review process or guiding a professional facilitator if one is used.
- Ensuring that the review is focused on substance, and
- Overseeing the production of the review report and signing off on the final report.

Composition of Panel

Because of the broad scope of the WAP Evaluation Study, a balance of technical expertise and subject area knowledge has to be established across the review panel. That balance is to include:

- *Technical expertise:* Experienced impact evaluators, experience with data collection/ measurement, and research design. The panel should include at least one statistician with advanced knowledge in econometrics.
- *Subject area knowledge:* Panel should have a balance of experts with experiences adequate to cover residential buildings, low-income energy programs, non-energy benefits, economics, and market assessment.

The WAP offers the following candidates for reviewers to the Chairperson (See attachment), although the Chairperson has responsibility to make final selection of reviewers. The Chairperson may wish to consult the list of possible candidates as necessary.

Process and Meeting Format

- Materials will be sent to reviewers [DATE]

- Review Chairman and Reviewers provide individual initial written comments and individually ratings to the WAP Manager and WAP Evaluation Design Team [DATE]

- A full day session (8:30 am – 5:00 pm) with 6 reviewers will be held in Washington DC [TENTATIVE DATE].
  - The external review panel will present their current thinking on the draft Evaluation Plan
  - The WAP Evaluation Design Team will respond to initial reviewer comments
  - A rigorous question & answer session moderated by the Review Chairman will follow
  - The panel will be given an opportunity to modify their initial individual ratings as necessary
  - The panel will provide a brief summary of their comments and recommendations and submit their final initial individual ratings to the WAP Manager and WAP Evaluation Design Team before adjourning.

- The Review Chairperson will develop the review report.

- Within one week following the review the Review Chairperson will summarize the reviewers' comments, ratings and recommendations.
- The WAP Evaluation Design Team will write a response to the review and formulate an action plan for revising the Plan.

Criteria and Questions for Judging the WAP Draft National Evaluation Plan

The Evaluation Plan is to be examined and reviewed on the basis of its technical quality. Inasmuch as credible evaluation findings reflect a soundness of design across the entire of the evaluation spectrum – from the design of the research through the data collection and analysis protocol to reporting – the external review panel is asked to assess and rate the current WAP Evaluation Plan on each of those criterion, using the scale presented below, in addition to written comments. In addition to rating the sections separately, each reviewer will be asked to give an overall rating for the Draft Evaluation Plan. Some aspects of technical quality (expressed as guiding questions) are provided below.

Research Design

- Are the research questions well formulated and relevant to the objectives of the evaluation?
- Are the metrics credible as measures of the outputs and outcomes being evaluated?
- Are the program and market logic models sound and credible?
- What is the research design?[47]
  - Is the choice of research design the most rigorous and feasible for answering the questions of the evaluation?
  - Are all applicable threats to the internal validity of the research design to be used identified, and methods for treatment articulated?
  - If the results are intended to be generalizable, are threats to external validity identified for treatment in the design?
- For statistical methods, were the degree of relationship between indicators, the tests of significance, and statistical precision for sample estimates built into the analysis and applied where possible?
- Does the research demonstrate understanding of previous related studies?

Data Collection

- Are the data collection and analysis methods credible, with appropriate quality assurance and control protocols?
- Are the measurement methods, instruments and their application credible?
- Are the assumptions guiding data collection valid?
- Are the data needs properly aligned to the evaluation research questions?
- Is the data collection protocol thorough – that is, are all the data required for meeting the explicit purposes of the evaluation planned for?
- If some missing data are to be inferred, is the proposed inference method appropriate?
- If a survey is planned, is potential non-response appropriately accounted for?

[47] To ensure the highest standard of performance in the review process, the expert reviewers are asked to become familiar with EERE's "Quality Assurance Guidance for Use of Impact Results Information," which provides guidance on classifying the strength of an evaluation study plan and draft report on the basis of the research design, treatment of threats to validity, and execution.

- Do the proposed data collection protocols provide fallback alternatives that are valid and reasonable for the scope and resources available?
- Are adequate plans for collecting the data and processing them (entry, cleaning and transforming as appropriate) provided?

Analysis

- Are the proposed analytic methods appropriate to the study designs, and are appropriate alternative analyses presented?
- Are the proposed analytic methods properly documented in sufficient detail to enable replication if necessary?

Reporting

- Is the report outline in compliance with DOE reporting guidelines?
- Would the final report provide the product needed for the uses for which the evaluation was initiated?

In addition to the above criterion and questions, the review panel is asked to also provide an overall assessment and rating of the draft technical evaluation study plan and the draft final report.

Materials Provided to Reviewers

- Brief background on the Weatherization Assistance Program
- Charter for the External Review
- WAP Draft Technical Evaluation Plan / WAP Draft Evaluation Report

Appendix H. Lessons Learned for Improving the Quality of EERE Evaluation Studies

Prepared by Ed Vine, LBNL

Introduction

A number of lessons have been learned from critiques of past EERE evaluation studies.[48] Awareness of these lessons can help promote continuous improvement in the planning, design, and conduct of evaluation studies in EERE. It is recommended that DOE evaluation project managers incorporate these lessons, as appropriate, into the statement of work used to hire an evaluation contractor, and use them to manage the evaluation study.

Formulation of the Request for Proposals (RFP)

1. ***Evaluation RFPs should be planned in the program's development-implementation cycle:*** Particularly for process evaluations, planning for the evaluations ideally should be concurrent with program planning, and the evaluation should be implemented as the program is being implemented. This will allow for timely data collection, the creation of needed databases as the program proceeds, more accurate data collection, and stronger evaluation designs to be implemented. For impact evaluations, the evaluation RFPs might be planned as the program is being planned, but the impact evaluation should wait until sufficient time has passed to allow program impacts to have occurred.

> Evaluation RFPs should be planned as the program is being planned. Process evaluations should be conducted as the program is being implemented, while impact evaluations should wait until sufficient time has passed so that program impacts have occurred.

2. ***Past evaluation experience should be clearly specified in the RFP:*** An RFP is typically open to any potential bidder. Since the objective of the evaluation study is to conduct a high quality and credible evaluation, it is critical that an offeror demonstrate past evaluation experience and writing skills. Most evaluators will have adequate knowledge of research designs, validity threats, survey sampling, and other technical topics, but actual relevant experience is an additional key selection consideration. While this may be obvious, there are cases where an offeror has good to excellent knowledge and experience but lacks demonstrated relevant experience for the particular evaluation subject. Likewise, an evaluator could lack strong writing skills which will complicate the project later in the final report preparation stage. Ask for a sample of an evaluator's Final Project Completion Report. In addition, the RFP should recognize that different professional evaluators have different specific expertise in the evaluation field. An evaluator who has experience in the evaluation of Research and Development (R&D) may not necessarily be the best choice for evaluating a technology deployment or market transformation program and vice versa.

[48] Much of this material is assembled from lessons learned by evaluators inside and outside of EERE and from summaries of comments made by external reviewers at evaluation study peer review meetings.

The RFP should clearly specify that a key priority in selecting a contractor would be demonstrated excellence in *relevant* past evaluation experience for the type of program being evaluated (e.g., R&D versus deployment) and strong writing skills.

3. ***Time for peer review should be clearly specified in the RFP:*** All EERE draft evaluation plans will be peer reviewed, since expert review is an excellent vehicle for ensuring a high quality and credible evaluation. Potential evaluation contractors should know that sufficient time has to be budgeted for the conduct of the peer review of the draft evaluation plan the draft evaluation report. This includes time to: participate in peer review meetings, respond to review comments, and revise the draft reports.

The RFP should clearly specify that sufficient time should be budgeted for conducting peer reviews of the draft evaluation plan and final report.

4. ***Identify specific responsibilities for members of the Evaluation Team***: There are different models for forming evaluation teams: (1) an evaluation expert is the Principal Investigator (PI) and has assembled people from the same organization for subject matter expertise; (2) an evaluation expert is the lead and subcontracts to other firms to add subject matter expertise; or (3) an evaluation expert is the PI and has sufficient subject matter expertise. EERE should not form the evaluation team, but allow the PI to form the evaluation team. When multiple team members are involved, it is important for the lead evaluator to indicate past working relationships with the other team members and to identify who is responsible for specific activities, such as: day to day project management, development of evaluation plan and statement of work, technical advice or execution of technology or market research content, research design, data collection, data analysis, report preparation, etc.

Specific responsibilities and past working relationships for Evaluation Team members need to be highlighted in the RFP and identified in the offeror's proposal that responds to the RFP.

Formulation of the Evaluation Statement of Work

5. ***Develop a Statement of Work (SOW) for the Evaluation Study:*** On occasion in the past, a program evaluation was initiated without preparing a full SOW. This often leads to an unproductive evaluation and wasted managerial time because a full consideration of the scope of the evaluation is not established *before hiring a contractor*.

Program staff should develop a preliminary SOW to use to hire an evaluation contractor. This will be part of the RFP. See Appendix A&B for a model statement of work and a specific example. Once the evaluator has been chosen, the preliminary SOW might be revised in preparing the evaluator's SOW.

6. ***Evaluation objective statements should be clearly specified in the evaluator's SOW:*** Evaluation SOWs do not always describe the intended uses of the evaluation, the unit of analysis, the decisions under consideration, the types of information required, or even clearly define the evaluation objectives. The evaluation should be designed with specific objectives in mind, and these should be clearly described in the SOW.

> Program staff initially, and then in consultation with the evaluation contractor, need to clarify intended uses of the evaluation, the unit of analysis, decisions under consideration, kinds of information required, and use this information to define clear evaluation objectives.

7. ***An evaluability assessment should be clearly specified in the evaluator's SOW:*** The first phase of an evaluation should be an Evaluability Assessment, which examines whether the project or program can be reliably and credibly evaluated. While Evaluability Assessments may be done at different levels of detail, the purpose of these is to determine the readiness of the project or program for evaluation, given the objectives of the evaluation, the intended uses of the results, the feasibility of obtaining the data, and the likelihood that the results are, in fact, verifiable in the time desired.

> The evaluation contractor should specify an Evaluability Assessment in the SOW.

Credibility of Results

8. ***Double counting:*** The overlapping and interactive structure of program components can lead to possible double counting of energy savings when savings estimates attributable to each program component (or activity) are developed separately. EERE deployment programs may use the outputs of EERE R&D programs. In such a case, both programs may claim credit for energy savings resulting from their efforts.

> For impact and cost-benefit evaluations, evaluation contractors should be asked to identify areas where double counting is possible and describe how double counting would be avoided, addressed, and documented in the report.

9. ***Sources of overestimation & underestimation:*** Often, impact evaluation studies report that their estimates are "conservative" in that overestimation is outweighed by underestimation. In other cases, spillover benefits from program outcomes may be hypothesized but not quantified because of the difficulty of making reliable estimates.

> For impact evaluations, evaluation contractors should be asked to clearly identify in the Evaluation Plan, and document in the report, all sources of overestimation & underestimation. Hypothesized spillover benefits should be discussed even if they are not quantified.

10. ***Use of "savings factors" in lieu of site-specific measurement:*** When savings factors, e.g., annual kWh saved per energy efficiency measure, are used in lieu of direct measurement, they must be applied appropriately to match the profile of the population that they are intended to represent. It generally will not be correct to transfer savings factors to entities that have widely different profiles compared to those from which the savings factors were derived.

> Evaluation contractors should be asked to fully describe the planned methodology for use of savings factors in the Evaluation Plan, including how they intend to account for site-by-site variation, applications variation, and other variations in the profile of the study population where these factors could be significant. Where savings factors are used, develop a means to check the reasonableness of the resultant energy savings numbers across the study population (e.g., acquire and evaluate information that can be used as a benchmark).

11. ***Identification of causal factors (potential confounders)*:** Evaluators need to consider and identify causal factors and all major rival explanations other than program activities that could contribute to observed changes in outcomes, and the possible implications of these factors on the outcomes. Given the multiple types of activities conducted by multiple organizations, it is important to identify possible causal factors and major rival explanations other than the program activities that are being evaluated.

> Evaluators need to consider and identify causal factors and all major rival explanations other than the program activities that could contribute to observed changes in outcomes in the Evaluation Plan and describe how they will be addressed in the analysis.

12. ***Construction of attribution questions in surveys*:** When survey-based questions are used as the basis to address attribution, the questions have to be carefully structured to get at the attribution issue at hand. Failure to properly structure the questions will result in unreliable recipient responses. For example, a question such as "Did it influence your decision—Yes or No?" is inadequate for addressing attribution. An attribution question should not force a "yes" or "no" response. Instead, it should distinguish response by degree of influence (e.g., very little, somewhat, significant, dominant; or a numeric degree-of-influence scale).

> Survey-based attribution questions must be properly constructed to capture the most valid estimate of the influence of the intervention, relative to the influence of other factors. Survey instruments should be reviewed by evaluation peers before the survey is fielded.

13. ***Construction of sustainability questions in surveys*:** When survey-based questions are used to address sustainability, the questions have to be carefully structured to get at the sustainability issue at hand. Failure to properly structure the questions will result in unreliable recipient responses. For example, sustainability questions need to differ for

program sustainability (e.g., whether a program is continuing without DOE funding) and market sustainability (e.g., whether a retrofit market is continuing without rebates), and by scale (local, state, regional or national levels). Sustainability issues that provide the context for these survey questions need to be addressed in the Evaluation Plan.

> Survey-based sustainability questions in draft survey instruments should allow for the many factors that can influence choice and be reviewed by evaluation peers before the survey is fielded.

14. ***Survey non-response:*** A common problem encountered in survey work is non-response. Non-response can introduce error into survey results. The degree to which the results represent the intended population critically depends on the response rate. A poor response rate can undermine the external validity of the survey results.

> Evaluation contractors who plan to use survey research should be asked to describe in the SOW and the Evaluation Plan their approach for avoiding, minimizing, or controlling potential non-response error. In the final report they should describe how they addressed non-response, and any implications for the reliability of the results. Evaluators should not consider the non-response problem for the first time *after* the survey is fielded.

15. ***Explicit documentation of the source(s) of energy savings:*** Sometimes studies that are not based on site measurement of savings fail to fully describe and document the source of their reported energy savings. Savings based on factors used by different sources, e.g., states, are provided without describing the assumptions underlying the savings factors.

> Evaluation contractors should explicitly address in the Evaluation Plan how they intend to estimate energy savings and the assumptions underlying their estimates. This should also be well documented in the final report.

16. ***Describing caveats on data used in the evaluation:*** Budget constraints sometimes force compromises in the methodology used for data collection, yet the potential weaknesses created by these necessary choices are not acknowledged. The study needs to be sure to fully describe the caveats and other issues concerning the study data and analysis (e.g., attribution, sample size limitations, elements of analysis, lack of validity testing, etc.).

> The report outline developed by the evaluation contractor should include a section on data and analysis limitations and caveats regarding the study findings. The report should adequately and appropriately highlight any concerns and limitations. Caveats should also be mentioned in the Executive Summary for the less reliable study findings and recommendations.

17. ***Sources of information:*** Some previous evaluation reports have not always described sources of data in sufficient detail and with transparency to allow an independent determination of the appropriateness of the information.

> The evaluation study SOW should stipulate that the evaluation contractor must describe sources of data in enough detail and with transparency to allow the appropriateness of the data to be determined. This description should be included in both the Evaluation Plan and the Final Report.

18. ***Potential biases:*** Previous evaluation reports have not always described potential biases that could be introduced in the collection and analysis of data. In the Evaluation Plan, evaluators need to discuss all potential biases and recommend approaches for controlling and correcting them.

> The Evaluation Plan needs to discuss all potential bases that could be introduced in the collection and analysis of data and recommend approaches for controlling and correcting them.

19. ***Logic modeling:*** Logic models are key elements of Evaluation Plans, and some evaluation plans and reports have not always provided good explanations and narratives of the logic models, leading to concerns the theory of change and the adequacy of metrics and evaluation questions. The logic models need to be structured so that the underlying program theory will be easily understood. This program theory would explain not only the issues that the program was created to address, but also identify the targets of the various program features, the expected changes, and the intended causal links regarding program processes.

> Structured and formal logic models need to be carefully described and explained in both the Evaluation Plan and the Final Report.

20. ***Baselines:*** The development of a credible baseline is a necessary element of an evaluation. In particular, the evaluator needs to describe not only what participants did before an intervention (e.g., receiving DOE funds) but also what non-participants were during this time period.

> Baselines need to be carefully described in both the Evaluation Plan and the Final Report.

21. ***Estimating leverage impact:*** For programs that provide funding to support investments in project areas, it is common for evaluation studies to attempt to estimate how much additional funding in project areas was leveraged by a program dollar (e.g., "one dollar in program dollars leveraged xx million dollars in additional funding"). These leverage impact estimates are sometimes grossly exaggerated. A common problem of evaluation studies when determining the leverage effect is that they do not always adequately address additionality or account for the nature of financing in the subject project areas. Estimates of leverage should be founded on a calculation method that addresses additionality (e.g., the timing of grantee and other investments, relative amounts of funding from sources, and importance of funds to the activity). The objective is to arrive at what is truly additional to EERE-provided funds for activities, using a valid leverage calculation methodology.

> For studies that attempt to estimate and report leverage impact, it is essential to determine the extent to which the "non-program players" (grantees and their partners) also devote dollars to activities in program-targeted project areas independently of program funding. Also, one must determine the amount of funds in the project areas that program beneficiaries would have invested even if the program funds were not available. Absent this and other information about the project financing, it will be difficult to know who is leveraging whom.

Interactions within Program and across Programs

22. ***Synergistic effects among program elements:*** Studies do not always make an effort to assess the synergistic effects among program elements – e.g., how a combination of publications, software tools, and technical assistance might be more effective than each as a separate entity.

> As appropriate, evaluation contractors should be asked to describe in the Evaluation Plan how they intend to assess the synergistic effects among program elements. However, avoid double counting. (See item #8)

23. ***The same population receives the services of multiple programs***. For example, how do deployment activities and other programs that provide direct service to the same set of customers interact to produce a customer choice? How should the resulting outcomes be allocated?

> Program staff should clearly document what other programs within or outside of EERE also serve their program's target audience. For impact evaluations, the Evaluation Plan should include a discussion of this issue and the plan for addressing it.

24. ***Accounting for "shelf life" of programs' products:*** The impacts of energy efficiency measures and practices most often continue to accrue beyond the retrospective period of an evaluation study. At the same time, these impacts do not last forever. In these

situations, it is appropriate to account for the effective useful life (EUL) of the technology or measure after installation/purchase. For both efficiency measures and deployed R&D technologies, there is an additional level of uncertainty introduced for benefits based on EUL. The effectiveness of most energy-efficient measures deteriorates with time. When conducting persistence studies for energy efficiency measures, careful attention must be paid to additional sources of bias and uncertainty introduced. Effective useful life and persistence should be included in reporting as additional impacts for a retrospective study, but there must be a distinction between benefits already achieved (i.e., realized), and the additional benefits from EUL and with consideration of persistence. i.e., the retrospective vs. lifetime impacts ought to be reported separately.

> EERE staff and the evaluation contractor should decide how to account for savings shelf life. The evaluation contractor should describe in the Evaluation Plan how this will be accomplished, and in a single report, separately report retrospective vs. lifetime impacts.

Findings and Recommendations Presented in Reports

25. ***Precision of reporting of the results:*** Studies sometimes report results at a level of precision that is not justified by the data and analysis.

> Evaluation contractors should not report numbers with too many decimal places. Also, in some cases, the evaluation contractor might consider reporting results as a point estimate within a range.

26. ***Provide a list of clear, actionable and prioritized recommendations that are supported by the analysis:*** Some evaluation studies have not developed program-improvement recommendations for the client to consider, or do not always develop recommendations that are adequately supported by the analysis. Similarly, recommendations for improving the quality of the evaluation are often omitted, even though the evaluation report acknowledges difficulties in performing the evaluation.

> Evaluation contractors should be asked to provide an explicit set of recommendations for both program and evaluation improvement, as appropriate, and ensure they are supported by the analysis conducted. Recommendations should be ranked in priority order (high, medium, low).

27. ***Rank findings by level of defensibility***: Outcome and impact evaluations that estimate impacts by program component or activity typically do not associate a level of defensibility to each reported component result.

For evaluations that use different methods to estimate impacts for program component or activities, evaluation contractors should report on the level of defensibility of each estimate associated with a particular program component/activity for which a quantified finding was developed. This need not be a quantitative value; a subjective ranking should be feasible based on the relative strengths and weaknesses of the respective method. This could be in addition to describing caveats for the findings.

28. ***Program record keeping and database recommendations:*** Program record keeping and databases are rarely designed to support evaluation activity. Often information about participants that is important for evaluation procedures is missing from program records.

Evaluation contractors should make explicit recommendations for routine program record-keeping and data collection, so the program can begin to collect data needed for future similar evaluations.

Appendix I. Example of a Technical Evaluation Plan Outline

1. **Introduction**

2. **Program Background**
    a. Program Objectives and Program Effects
    b. Program Technologies and Services

3. **Research Objectives**
    a. Objectives
    b. Research Questions

4. **Data Requirements and Acquisition**
    a. Data Requirements
    b. Data Acquisition
    c. Data Cleaning and Quality Control

5. **Sampling Approach**
    a. Program Status
    b. Sample Frame
    c. Design M&V Sample
    d. Design Billing Analysis Sample

6. **Evaluation Approach**
    a. Overall Scope of the Evaluation
    b. Detailed Evaluation Activities
        i. Conduct File Reviews
        ii. Design Survey and Data Collection Forms
        iii. Establish Baseline Condition
        iv. Conduct Telephone Surveys and/or On Site Inspections
        v. Gross Impact Methodology
    c. Billing Analysis Approach
        i. Regression Approach
    d. Calculation of Additional Metrics
        i. Lifetime Energy Savings
        ii. Greenhouse Gas Emission Savings
        iii. Demand Savings
        iv. Cost Savings
    e. Net-to-Gross Methodology
        i. Free-Ridership
        ii. Participant Spillover
        iii. Non-participant Spillover
        iv. Net-to-Gross Ratios
    f. Extrapolate Sample Results to the Population
        i. M&V Sample Extrapolation
        ii. Billing Analysis Extrapolation
        iii. Overall Program Extrapolation
        iv. Dealing with Inaccurate or Inconsistent Reported Savings

g. Economic Impacts Analysis

7. Management Plan

a. Project Management
b. Reporting
c. Schedule and Milestones

8. References

9. Appendices

Appendix J. American Evaluation Association Ethical Principles for Evaluators[49]

Guiding Principles for Evaluators[50]

Revisions reflected herein ratified by the AEA membership, July 2004

Preface: Assumptions Concerning Development of Principles

A. Evaluation is a profession composed of persons with varying interests, potentially encompassing but not limited to the evaluation of programs, products, personnel, policy, performance, proposals, technology, research, theory, and even of evaluation itself. These principles are broadly intended to cover all kinds of evaluation. For external evaluations of public programs, they nearly always apply. However, it is impossible to write guiding principles that neatly fit every context in which evaluators work, and some evaluators will work in contexts in which following a guideline cannot be done for good reason. The Guiding Principles are not intended to constrain such evaluators when this is the case. However, such exceptions should be made for good reason (e.g., legal prohibitions against releasing information to stakeholders), and evaluators who find themselves in such contexts should consult colleagues about how to proceed.

B. Based on differences in training, experience, and work settings, the profession of evaluation encompasses diverse perceptions about the primary purpose of evaluation. These include but are not limited to the following: bettering products, personnel, programs, organizations, governments, consumers and the public interest; contributing to informed decision making and more enlightened change; precipitating needed change; empowering all stakeholders by collecting data from them and engaging them in the evaluation process; and experiencing the excitement of new insights. Despite that diversity, the common ground is that evaluators aspire to construct and provide the best possible information that might bear on the value of whatever is being evaluated. The principles are intended to foster that primary aim.

C. The principles are intended to guide the professional practice of evaluators, and to inform evaluation clients and the general public about the principles they can expect to be upheld by professional evaluators. Of course, no statement of principles can anticipate all situations that arise in the practice of evaluation. However, principles are not just guidelines for reaction when something goes wrong or when a dilemma is found. Rather, principles should proactively guide the behaviors of professionals in everyday practice.

D. The purpose of documenting guiding principles is to foster continuing development of the profession of evaluation, and the socialization of its members. The principles are meant to stimulate discussion about the proper practice and use of evaluation among members of the profession, sponsors of evaluation, and others interested in evaluation.

E. The five principles proposed in this document are not independent, but overlap in many ways. Conversely, sometimes these principles will conflict, so that evaluators will have to choose among them. At such times evaluators must use their own values and knowledge of the setting to determine the appropriate response. Whenever a course of action is unclear, evaluators should solicit the advice of fellow evaluators about how to resolve the problem before deciding how to proceed.

---

[49] The American Evaluation Association developed these ethical principles to guide the professional practice of evaluators. EERE expects evaluation experts who perform general program evaluation of its programs to be governed by these principles.

[50] The American Evaluation Association provides these Guiding Principles online at www.eval.org/Publications/GuidingPrinciples.asp. Last accessed: 4/29/14.

F. These principles are intended to supersede any previous work on standards, principles, or ethics adopted by AEA or its two predecessor organizations, the Evaluation Research Society and the Evaluation Network. These principles are the official position of AEA on these matters.
G. These principles are not intended to replace standards supported by evaluators or by the other disciplines in which evaluators participate.
H. Each principle is illustrated by a number of statements to amplify the meaning of the overarching principle, and to provide guidance for its application. These illustrations are not meant to include all possible applications of that principle, nor to be viewed as rules that provide the basis for sanctioning violators.
I. These principles were developed in the context of Western cultures, particularly the United States, and so may reflect the experiences of that context. The relevance of these principles may vary across other cultures, and across subcultures within the United States.
J. These principles are part of an evolving process of self-examination by the profession, and should be revisited on a regular basis. Mechanisms might include officially-sponsored reviews of principles at annual meetings, and other forums for harvesting experience with the principles and their application. On a regular basis, but at least every five years, these principles ought to be examined for possible review and revision. In order to maintain association-wide awareness and relevance, all AEA members are encouraged to participate in this process.

The Principles

A. Systematic Inquiry: Evaluators conduct systematic, data-based inquiries, and thus should:

1. Adhere to the highest technical standards appropriate to the methods they use.
2. Explore with the client the shortcomings and strengths of evaluation questions and approaches.
3. Communicate the approaches, methods, and limitations of the evaluation accurately and in sufficient detail to allow others to understand, interpret, and critique their work.

B. Competence: Evaluators provide competent performance to stakeholders, and thus should:

1. Ensure that the evaluation team collectively possesses the education, abilities, skills, and experience appropriate to the evaluation.
2. Ensure that the evaluation team collectively demonstrates cultural competence and uses appropriate evaluation strategies and skills to work with culturally different groups.
3. Practice within the limits of their competence, decline to conduct evaluations that fall substantially outside those limits, and make clear any limitations on the evaluation that might result if declining is not feasible.
4. Seek to maintain and improve their competencies in order to provide the highest level of performance in their evaluations.

C. Integrity/Honesty: Evaluators display honesty and integrity in their own behavior, and attempt to ensure the honesty and integrity of the entire evaluation process, and thus should:

1. Negotiate honestly with clients and relevant stakeholders concerning the costs, tasks, limitations of methodology, scope of results, and uses of data.
2. Disclose any roles or relationships that might pose a real or apparent conflict of interest prior to accepting an assignment.
3. Record and report all changes to the original negotiated project plans, and the reasons for them, including any possible impacts that could result.
4. Be explicit about their own, their clients', and other stakeholders' interests and values

related to the evaluation.

5. Represent accurately their procedures, data, and findings, and attempt to prevent or correct misuse of their work by others.
6. Work to resolve any concerns related to procedures or activities likely to produce misleading evaluative information, decline to conduct the evaluation if concerns cannot be resolved, and consult colleagues or relevant stakeholders about other ways to proceed if declining is not feasible.
7. Disclose all sources of financial support for an evaluation, and the source of the request for the evaluation.

D. Respect for People: Evaluators respect the security, dignity and self-worth of respondents, program participants, clients, and other evaluation stakeholders, and thus should:

1. Seek a comprehensive understanding of the contextual elements of the evaluation.
2. Abide by current professional ethics, standards, and regulations regarding confidentiality, informed consent, and potential risks or harms to participants.
3. Seek to maximize the benefits and reduce any unnecessary harm that might occur from an evaluation and carefully judge when the benefits from the evaluation or procedure should be foregone because of potential risks.
4. Conduct the evaluation and communicate its results in a way that respects stakeholders' dignity and self-worth.
5. Foster social equity in evaluation, when feasible, so that those who give to the evaluation may benefit in return.
6. Understand, respect, and take into account differences among stakeholders such as culture, religion, disability, age, sexual orientation and ethnicity.

E. Responsibilities for General and Public Welfare: Evaluators articulate and take into account the diversity of general and public interests and values, and thus should:

1. Include relevant perspectives and interests of the full range of stakeholders.
2. Consider not only immediate operations and outcomes of the evaluation, but also the broad assumptions, implications and potential side effects.
3. Allow stakeholders' access to, and actively disseminate, evaluative information, and present evaluation results in understandable forms that respect people and honor promises of confidentiality.
4. Maintain a balance between client and other stakeholder needs and interests.
5. Take into account the public interest and good, going beyond analysis of particular stakeholder interests to consider the welfare of society as a whole.

Background

In 1986, the Evaluation Network (ENet) and the Evaluation Research Society (ERS) merged to create the American Evaluation Association. ERS had previously adopted a set of standards for program evaluation (published in New Directions for Program Evaluation in 1982); and both organizations had lent support to work of other organizations about evaluation guidelines. However, none of these standards or guidelines were officially adopted by AEA, nor were any other ethics, standards, or guiding principles put into place. Over the ensuing years, the need for such guiding principles was discussed by both the AEA Board and the AEA membership. Under the presidency of David Cordray in 1992, the AEA Board appointed a temporary committee chaired by Peter Rossi to examine whether AEA should address this matter in more detail. That committee issued a report to the AEA Board on November 4, 1992, recommending that AEA should pursue this matter further. The Board followed that recommendation, and on that date

created a Task Force to develop a draft of guiding principles for evaluators. The task force members were:

William Shadish, Memphis State University (Chair)
Dianna Newman, University of Albany/SUNY
Mary Ann Scheirer, Private Practice
Chris Wye, National Academy of Public Administration

The AEA Board specifically instructed the Task Force to develop general guiding principles rather than specific standards of practice. Their report, issued in 1994, summarized the Task Force's response to the charge.

Process of Development. Task Force members reviewed relevant documents from other professional societies, and then independently prepared and circulated drafts of material for use in this report. Initial and subsequent drafts (compiled by the Task Force chair) were discussed during conference calls, with revisions occurring after each call. Progress reports were presented at every AEA board meeting during 1993. In addition, a draft of the guidelines was mailed to all AEA members in September 1993 requesting feedback; and three symposia at the 1993 AEA annual conference were used to discuss and obtain further feedback. The Task Force considered all this feedback in a December 1993 conference call, and prepared a final draft in January 1994. This draft was presented and approved for membership vote at the January 1994 AEA board meeting.

Resulting Principles. Given the diversity of interests and employment settings represented on the Task Force, it is noteworthy that Task Force members reached substantial agreement about the following five principles. The order of these principles does not imply priority among them; priority will vary by situation and evaluator role.

A. **Systematic Inquiry:** Evaluators conduct systematic, data-based inquiries about whatever is being evaluated.

B. **Competence:** Evaluators provide competent performance to stakeholders.

C. **Integrity/Honesty:** Evaluators ensure the honesty and integrity of the entire evaluation process.

D. **Respect for People:** Evaluators respect the security, dignity and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.

E. **Responsibilities for General and Public Welfare:** Evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.

Recommendation for Continued Work. The Task Force also recommended that the AEA Board establish and support a mechanism for the continued development and dissemination of the Guiding Principles, to include formal reviews at least every five years. The Principles were reviewed in 1999 through an EvalTalk survey, a panel review, and a comparison to the ethical principles of the Canadian and Australasian Evaluation Societies. The 2000 Board affirmed this work and expanded dissemination of the Principles; however, the document was left unchanged.

Process of the 2002-2003 Review and Revision. In January 2002 the AEA Board charged its standing Ethics Committee with developing and implementing a process for reviewing the Guiding Principles that would give AEA's full membership multiple opportunities for comment. At its Spring 2002 meeting, the AEA Board approved the process, carried out during the ensuing months. It consisted of an online survey of the membership that drew 413 responses, a "Town Meeting" attended by approximately 40 members at the Evaluation 2002 Conference, and a compilation of stories about evaluators' experiences relative to ethical concerns told by AEA members and drawn from the *American Journal of Evaluation*. Detailed findings of all three sources of input were reported to the AEA Board in *A Review of AEA's Guiding Principles for Evaluators*, submitted January 18, 2003.

In 2003 the Ethics Committee continued to welcome input and specifically solicited it from AEA's Diversity Committee, Building Diversity Initiative, and Multi-Ethnic Issues Topical Interest Group. The first revision reflected the Committee's consensus response to the sum of member input throughout 2002 and 2003. It was submitted to AEA's past presidents, current board members, and the original framers of the Guiding Principles for comment. Twelve reviews were received and incorporated into a second revision, presented at the 2003 annual conference. Consensus opinions of approximately 25 members attending a Town Meeting are reflected in this, the third and final revision that was approved by the Board in February 2004 for submission to the membership for ratification. The revisions were ratified by the membership in July of 2004.

The 2002 Ethics Committee members were:
Doris Redfield, Appalachia Educational Laboratory (Chair)
Deborah Bonnet, Lumina Foundation for Education
Katherine Ryan, University of Illinois at Urbana-Champaign
Anna Madison, University of Massachusetts, Boston

In 2003 the membership was expanded for the duration of the revision process:
Deborah Bonnet, Lumina Foundation for Education (Chair)
Doris Redfield, Appalachia Educational Laboratory
Katherine Ryan, University of Illinois at Urbana-Champaign
Gail Barrington, Barrington Research Group, Inc.
Elmima Johnson, National Science Foundation

Appendix K. Program Evaluation Glossary

Evaluation and performance measurement professionals use terms that are common to the field of program evaluation. Knowledge of how these terms are used by evaluators will help program managers communicate their evaluation needs and expectations to a contractor. The definitions are listed alphabetically.

The terminology in this glossary reflects usage by EERE programs and federal evaluations experts. For example, although energy savings terminology appears in the glossary, the goals of EERE programs and the corresponding evaluation objectives are not limited to saving energy—although that may be the long-term outcome sought. R&D programs may focus on producing new materials or products. Some programs focus on training and creating infrastructure for which energy savings goals are not established. The definitions used are often more comprehensive to reflect these broader goals.

Accuracy. The degree of correspondence between the measurement made on an indicator and the true value of the indicator at the time of measurement.

Activities. The action steps necessary to produce program outputs.

Additionality. A criterion applied to a non-energy outcome, e.g., greenhouse gas emissions (GHG), stipulating that the non-energy benefit should only be quantified if the benefit would not have happened anyway. This criterion has been used R&D program evaluations in which a DOE program has contributed to a measureable increase in a non-energy benefit, but part of that increase should be attributed to another agency or the private sector which turned the research into a beneficial outcome. This amounts to removing, or "netting out," of the total outcome for a program the proportion of the outcome due to another entity's activity but still giving DOE credit for its contribution to the outcome.

Attribution. Ascribing or establishing a causal relationship between action(s) taken by a program and an outcome.

Baseline. Conditions, including energy consumption and related non-energy benefits such as emissions reductions and jobs that would have occurred in a population served by a program without implementation of the program. Baselines can be defined as either program-specific baselines or performance standard baselines, e.g., building codes. See also Counterfactual.

Benchmark. A measurement or standard that serves as a point of reference by which process performance is measured.

Benefit-Cost Ratio. The mathematical relationship between the benefits and costs associated with the implementation of a program. The benefits and costs are typically expressed in dollars.

Bias. The extent to which a measurement or a sampling or an analytic method systematically underestimates or overestimates a value.

Billing Analysis. An analytic methodology used to estimate program effects based on the use of the energy consumption data contained in consumer billing data. Two billing analysis methods are available. The first method compares the billing data from program participants over a period

of time (usually a year) before the energy-related actions were taken to the participants' billing data for comparable periods of time after the actions were taken. The second method compares the difference between the before-and-after data for the group of participants to the before-and-after difference between a comparable group of non-participants to estimate the difference of the differences.

Building Energy Simulation Model. A computer model based on physical engineering principals and/or standards used to estimate energy usage and/or savings. These models usually incorporate site-specific data on customers and physical systems such as square footage, weather, surface orientations, elevations, space volumes, construction materials, equipment use, lighting and building occupancy. Building simulation models can usually account for interactive effects between end uses, part-load efficiencies, and changes in external and internal heat gains/losses.

Confidence. An indication of the a probability that the true value of the quantity in question lies within a specified distance of the estimated value of the value, as developed by a sample. The distance is also called the sampling error or interval.

Confidence Interval. A specification consisting of the probability that the true value for an estimate developed by a sample (see Confidence) lies within a certain range or interval around the estimate and the range or interval itself (see Precision). Confidence interval is often expressed as +/-10% with 90% confidence, or often, 90 +/-10%.

Comparison Group. A group of individuals or organizations that have not had the opportunity to receive program benefits that is measured to determine the extent to which its members have taken actions promoted by the program. Like a control group, the comparison group is used to measure the level to which the promoted actions would have been taken if the program did not exist. However, unlike a control group, a comparison group is chosen through methods other than randomization, e.g., selection on the basis of similar demographic characteristics. See also Representative Sample.

Construct. An attribute, usually unobservable, such as attitude or comfort, that is represented by an observable measure.

Control Group. A randomly selected group of individuals or organizations that have not had the opportunity to receive program benefits that is measured to determine the extent to which its members have taken actions promoted by the program. The control group is used to measure the level to which the promoted actions would have been taken if the program did not exist. See also Comparison Group.

Correlation. For a set of observations, such as for participants in a program, the extent to which high values for one variable are associated with high values of another variables for the same participant. For example, facility size and energy consumption usually have a high positive correlation.

Cost-Benefit and Cost-Effectiveness. Comparison of a program's outputs or outcomes with the costs (resources expended) to produce them. Cost-effectiveness Evaluation analysis assesses the cost of meeting a single goal or objective, and can be used to identify the least costly alternative to meet that goal. Cost-benefit analysis aims to identify and compare all relevant costs and

benefits, usually expressed in dollar terms. The two terms are often interchanged in evaluation discussions.

Cross-Sectional Data. Observations collected on subjects or events at a single point in time.

Counterfactual. The amount of a program's measured outcome that would have occurred had the program never occurred. The counterfactual is an unmeasureable construct of how the present in the absence of the program's influence. See Baseline.

Deemed Savings. An estimate of an energy savings or energy-demand savings outcome (gross savings) for a single unit of an installed energy-efficiency or renewable-energy measure that (1) has been developed from data sources and analytical methods that are widely considered acceptable for the measure and purpose, and (2) will be applied to situations other than that for which it was developed. That is, the unit savings estimate is "deemed" to be acceptable for other applications. Deemed savings estimates are more often used in program planning than in evaluation. They should not be used for evaluation purposes when a program-specific evaluation can be performed. When a deemed savings estimate is used, it is important to know whether its baseline is an energy-efficiency code or open-market practice. The most extensive database of deemed savings is California's Database for Energy Efficiency Resources (DEER). The deemed savings in DEER are tailored to California. Deemed savings are sometimes termed "stipulated savings."

Defensibility. The ability of evaluation results to stand up to scientific criticism. Defensibility is based on assessments by experts of the evaluation's validity, reliability, and accuracy. See also Rigor and Strength.

Direct Customers. The individuals or organizations that receive the outputs of a program.

Experimental Design. A method of estimating the amount of an outcome attributable to a program or other event in which outcomes between at least two randomly assigned groups are compared.

External Factor. A factor that may enhance or nullify underlying program assumptions and thus the likelihood of goal achievement. Goal achievement may also be predicated on certain conditions (events) not happening. They are introduced by external forces or parties, and are not of the agency's own making. The factors may be economic, demographic, social, or environmental, and they may remain stable, change within predicted rates, or vary to an unexpected degree.

External Validity. The extent to which a finding applies (or can be generalized) to persons, objects, settings, or times other than those that were the subject of study.

Evaluation. Evaluations are systematic, objective studies conducted periodically or on an ad hoc basis to assess how well a program is working. They help managers determine if timely adjustments are needed in program design to improve the rate, or quality, of achievement relative to the committed resources

Generalizability. Used interchangeably with "external validity."

Gross Savings. The change in energy consumption and/or demand that results directly from program-related actions taken by participants in an energy efficiency program, regardless of other potential causes of the observed change. The use of the adjective “gross” when applied to any energy program outcome means the measured or estimated outcome has not been adjusted for influences other than the program which may have been responsible for some portion of the outcome.

Independent Variables. The factors that affect a programs outcome(s) that cannot be controlled, e.g., weather, historical events.

Impact Evaluation. The application of scientific research methods to estimate how much of the observed results, intended or not, are caused by program activities and how much might have been observed in the absence of the program. This form of evaluation is employed when external factors are known to influence the program's outcomes in order to isolate the program's contribution to achievement of its objectives.

Indicator (also Performance Indicator). A particular characteristic used to measure outputs or outcomes; a quantifiable expression used to observe and track the status of a process. An indicator constitutes the observable evidence of accomplishments, changes made, or progress achieved.

Internal Validity. The extent to which the causes of an effect are established by an inquiry.

International Performance Measurement and Verification Protocol (IPMVP). A guidance document with a framework and definitions describing four approaches to measurement and verification. The document is published and maintained by the Efficiency Valuation Organization (www.evo-world.org)

Logic Model. A plausible and sensible diagram of the sequence of causes (resources, activities, and outputs) that produce the effects (outcomes) sought by the program.

Longitudinal Data. Observations collected over a period of time. The sample (instances or cases) may or may not be the same each time but the population remains constant. Longitudinal data are sometimes called "time series data."

Measurement. A procedure for assigning a number to an observed object or event.

Measurement and Verification (M&V). A set of practices in program performance monitoring and evaluation in that is associated with the documentation of outcomes at individual participant sites using one or more methods that can involve direct measurements, engineering calculations, statistical analyses, and/or computer simulation monitoring. The IPVMP defines four standard M&V approaches. M&V does not estimate attribution of the measured outcomes.

Monitoring. The collection of relevant measurement data over time at a facility, e.g., energy and water consumption, for the purpose of outcome analysis or observing system performance.

Needs/Market Assessment Evaluation. An evaluation that assesses market baselines, customer needs, target markets, and how best to address these issues by the program in question. Findings help managers decide who constitutes the program's key markets and clients and how to best

serve the intended customers. When performed at the beginning of a program, needs/market assessment evaluations also establish baselines against which to compare future progress.

Net Savings. The change in energy consumption and/or demand that is attributable to a particular program. The use of the adjective “net” when applied to any energy program outcome means the measured or estimated outcome has been adjusted for influences other than the program which may have been responsible for some portion of the outcome.

Non-Energy Benefits. The identifiable non-energy effects associated with program implementation or participation. Examples include avoided greenhouse gas emissions, productivity improvements, and job creation.

Outcome. Changes or benefits resulting from activities and outputs. Programs typically have multiple, sequential outcomes, sometimes called the program's outcome structure. First, there are "short term outcomes", those changes or benefits that are most closely associated with or "caused" by the program's outputs. Second, there are "intermediate outcomes," those changes that result from an application of the short-term outcomes. "Longer term outcomes" or program impacts, follow from the benefits accrued though the intermediate outcomes.

Output. The product, good, or service offered to a program's direct customers.

Panel Data. A special form of longitudinal data in which observations are collected on the same sample of respondents over a period of time.

Peer Review. Objective review and advice from peers. EERE defines peer review as: "A rigorous, formal, and documented evaluation process using objective criteria and qualified and independent reviewers to make a judgment of the technical/ scientific/business merit, the actual or anticipated results, and the productivity and management effectiveness of programs and/or projects."

Performance Measure. An indicator, statistic or metric used to gauge program performance. Also referred to as a performance indicator.

Performance Measurement. The process of developing measurable indicators that can be systematically tracked to assess progress made in achieving predetermined goals and using such indicators to assess progress in achieving these goals.

Persistence. The estimated or described changes in net program impacts over time taking into consideration all known factors that degrade the performance of a desired outcome, including retention in use and technical degradation of equipment performance.

Precision. The band or interval of values around an estimate of an outcome, as developed from a sample, within which the true value of the outcome lies. Precision is expressed as +/- some value. The interval may be expressed as a percentage or an absolute value. In combination with the “confidence” of the interval, the evaluator can specify the probability that the true value lies within this interval (see Confidence Interval). For a given probability, the smaller the interval, the more precise the estimate. Precision is also used to represent the degree to which an estimated result would be replicated with repeated studies conducted using the same research design.

Probability Sampling. A method for drawing a sample from a population such that all possible samples have a known and specified probability of being drawn.

Portfolio/Program Portfolio. Either (1) a collection of similar programs addressing the same market (e.g., a portfolio of residential programs), technology (e.g., motor efficiency programs), or mechanisms (e.g., loan programs), or activities (e.g., R&D programs), or (2) the set of all programs conducted by one organization (e.g., EERE).

Precision. The closeness of agreement among repeated measurements of the same physical quantity. The term precision is used in evaluation as an indicator of how close a program outcome that has been estimated from a sample is to the true impact of a program. Precision may expressed as an interval around an estimate within which the true value lies. The interval may be expressed in absolute values or as a percent of the estimate. (See also Confidence Interval.)

Process (or Implementation Process). Assessment of the extent to which a program is operating as intended. Process evaluation examines the efficiency and effectiveness of program implementation processes. It assesses program activities' conformance to program design, to professional standards or customer expectations, and to statutory and regulatory requirements.

Program. "Program" refers to a collection of activities that are unified with respect to management structure and overall goal.

Program Theory. A presentation of the goals of a program, incorporated with a detailed presentation of the activities that the program will use to accomplish those goals and the identification of the causal relationships between the activities and the program's outcomes. Program theory is often the basis for a logic model.

Portfolio. A collection of projects. A single individual or organization can have multiple R&D portfolios.

Qualitative Data. Information expressed in the form of words.

Quantitative Data. Information expressed in the form of numbers. Measurement gives a procedure for assigning numbers to observations. See Measurement.

Random Assignment. A method for assigning subjects to one or more groups by chance.

Randomized Control Trial (RCT). A type of experimental program design in which objects in the program's target field, or audience, are randomly assigned to two groups: a treatment group that receives the programs' outputs (participants) and a control group that does not (non-participants). The program's subsequent experimental evaluation design compares the outcomes for these two groups to develop an estimate of the outcomes attributable to the program.

Regression Analysis. A statistical analysis method that quantifies the mathematical relationship between a dependent variable (response variable) and specified independent variables (explanatory variables). The mathematical model of their relationship is called a regression equation.

Replication. An outcome effect that occurs when energy savings identified at a site are implemented elsewhere, e.g., at a different site, internal or external to the site. The replication process usually is initiated at a program-participant site. (See also Spillover.)

Reliability. The quality of a measurement process that would produce similar results on: (1) repeated observations of the same condition or event; or (2) multiple observations of the same condition or event by different observers.

Representative Sample. A sample that has approximately the same distribution of characteristics as the population from which it was drawn.

Resources. Human and financial inputs as well as other inputs required to support the program's activities.

Retention. An outcome effect that describes the degree to which measures or practices are retained in use after they are installed or implemented.

Rigor. A subjective term describing the degree of confidence that one can have that the results of the evaluation correctly describe the actual results, so that, if the evaluation were repeated many times, it would produce the same results (see Reliability). The many components of an evaluation activity, e.g., statistical confidence interval, measurement accuracy, treatment of missing measurements, the research design, and adherence to the evaluation plan, contribute to a judgment of the rigor of the evaluation. A high level of rigor makes it easier to defend the results (see Defensibility).

Sample. A portion of the population served by a program selected to represent the whole population served by a program. Differing evaluation approaches rely on simple, stratified, or representative samples.

Sample Design. The specification of the approach used to select sample units.

Sampling Error. An error that arises because the data are collected from a part, rather than the whole of the population served by a program. It is usually quantifiable from the sample data in the case of probability sampling. (See Confidence Interval.)

Simple Random Sample. A method for drawing a sample from a population such that all samples of a given size have equal probability of being drawn.

Significance Level. The probability of getting a particular value in a sample result, e.g., a mean of 43.0, or a proportion of 0.6, or a difference between two means of 3.0, or a quantitative relationship between the program treatment and an outcome—when, in fact, the hypothesized true value is some other value (that you must specify beforehand, e.g., a zero difference). The probability is often expressed using the Greek letter alpha (a) and should also be chosen before the data are collected. Probabilities (significance levels) of less than 0.1, 0.05, or 0.01 are typically selected for tests of significance.

Spin-off. Savings estimates that are based on verbal or undocumented recommendations from an energy-efficiency program output.

Spillover. The benefit of a program intervention that accrues to individuals or organizations that are not direct recipients of the program's outputs.

Stratified Sample. A sampling method where the population is divided into X units of subpopulations, called strata, that are non-overlapping and together make up the entire population. When random sampling is used, a simple random sample is taken of each stratum to create the total sample.

Strength. A term used to describe the overall defensibility of the evaluation as assessed by use of scientific practice, asking appropriate evaluation questions, documenting assumptions, making accurate measurements, and ruling out competing evidence of causation.

Structured Interview. An interview in which the questions to be asked, their sequence, and the detailed information to be gathered are all predetermined; used where maximum consistency across interviews and interviewees is needed.

Treatment Group. The subjects of the intervention being studied. See also Direct Customers.

Triangulation. An evaluation approach that compares the results from two or more different data collection, measurement, or analytical techniques on the same problem or goal set to derive a "best" estimate from the analysis of the comparison.

Uncertainty. Usually interpreted as the range or interval of doubt surrounding a measured or calculated value within which the true value is expected to fall with some degree of confidence. See Confidence Interval. When this concept includes consideration of measurement accuracy and threats to the validity of an evaluation research design, it becomes a component of rigor. See Rigor.

Validity. See Internal Validity and External Validity.

Verification. An independent assessment that the program has been implemented in accordance with the program design. Verification activities are generally conducted during on-site surveys of a sample of program activities. It may include one-time or multiple activities over the life of the program. Verification is a subset of evaluation and, as such, can also include review and confirmation of evaluation methods used, samples drawn, and calculations used to estimate program outcomes and impacts.

DOE/EE-1227 • August 2015

Printed with a renewable-source ink on paper containing at least
50% wastepaper, including 10% post-consumer waste.

U.S. Department of Energy
Energy Efficiency and Renewable Energy
Bringing you a prosperous future where energy is clean, abundant, reliable, and affordable